DOCUMENT RESUME

ED 082 754 LI 004 474

TITLE	Clustering of Journal Titles According to Citation Data: Report on Preparatory Work, Design, Data Collection, and Preliminary Analyses. Design of Information Systems in the Social Sciences, Working Paper No. 11.
INSTITUTION	Bath Univ. of Technology (England). Univ. Library.
SPONS AGENCY	Office for Scientific and Technical Information, London (England).
PUB DATE	Sep 73
NOTE	101p.; (50 references)
EDRS PRICE	MF-$0.65 HC-$6.58
DESCRIPTORS	Bibliographic Citations; *Cluster Analysis; *Cluster Grouping; Information Systems; Periodicals; *Scholarly Journals; Serials; *Social Sciences
IDENTIFIERS	*Titles

ABSTRACT

DISISS (Design of Information Systems in the Social Sciences) is a research project financed by OSTI, which began in January 1971. The objective of the project is to carry out research necessary for the effective design of information systems in the social sciences. The aim of this part of the DISISS project is the application of statistical techniques to citation data in order to group journal titles in the social sciences. Various statistical techniques exist, some with a fairly long history, for grouping items according to observable attributes. Details of these techniques and the selection of one for use with DISISS data are discussed. This report coveres the use of cluster techniques in bibliography, techniques of clustering, an analysis of the pilot study data, progress with data collection and conversion, and work that is required for the future. (Other reports in the DISISS series are ED 060876, 072815, 072816 and LI 004 401 through 004 403.)
(Author/SJ)

ED 082754

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
Bath University Library
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER.

Bath University Library

Design of Information Systems in the Social Sciences

Working Paper No. 11

CLUSTERING OF JOURNAL TITLES ACCORDING TO CITATION DATA: REPORT ON PREPARATORY WORK, DESIGN, DATA COLLECTION, AND PRELIMINARY ANALYSES

September 1973

FILMED FROM BEST AVAILABLE COPY

LI 004 474

CONTENTS

		Page No.
PREFACE		2
1	INTRODUCTION	3
	1.1 Statement of aims	3
	1.2 Data to be used	3
	1.3 Schedule of work	3
	1.4 Reasons for using citation data	4
2	PREVIOUS WORK	6
3	USE OF CLUSTER TECHNIQUES IN BIBLIOGRAPHY	9
	3.1 Implications of clustering for the design of information systems	9
4	TECHNIQUES OF CLUSTERING	11
	4.1 Clustering terms	11
	4.2 Clustering techniques	11
	4.2.1 The SCICON algorithm	13
	4.2.2 Reasons for selection of the SCICON algorithm	17
	4.2.3 Comparison with other approaches and algorithms	19
	4.3 Data for clustering	31
	4.4 Treatment of data before clustering	33
	4.4.1 Reduction in size of the data matrix	34
	4.4.2 Normalization, self-citation adjustment and scaling	35
	4.4.3 Choice of data adjustments	37
	4.5 Reliability and stability of clusters	37
	4.6 Evaluation and representation of clusters	39
5	ANALYSIS OF THE PILOT STUDY DATA	41
	5.1 Modification to clustering program for SCICON method	41
	5.2 Description of the pilot study data	42
	5.3 Results using the SCICON algorithm	42
	5.4 Results using other approaches	51
	5.5 Conclusions	55
6	PROGRESS WITH DATA COLLECTION AND CONVERSION	60
	6.1 ISI data	60
	6.2 Data collected in the field	62
	6.3 Conversion of field collected data	62
7	FUTURE WORK	

CONTENTS (cont'd)

		Page No.
REFERENCES		65
APPENDICES		
A	Source journals for the pilot study	70
B	Pilot study citation data	71
C	Source journals for the main study.	78
D	Criterion used in the SCICON algorithm for optimizing the division of the points into clusters	89
E	Clustering results using the SCICON algorithm	92
F	Clusters satisfying Van Rijsbergen's condition	98

PREFACE

DISISS (Design of Information Systems in the Social Sciences) is a research project financed by OSTI, which began in January 1971. The objective of the project is to carry out research necessary for the effective design of information systems in the social sciences. The project is based at the University of Bath and other organizations involved are the Polytechnic of North London and the Open University. Until May 1972 the University of Sussex was also involved.

The present working paper describes work on the clustering of journal titles by citation data. The work was carried out partly at the University of Sussex and partly at the Open University. The working paper describes work up to the end of April 1973. It was written by Mrs C.R. Arms and Mr W.Y. Arms with assistance from Mr J.M. Brittain and Mr S.A. Roberts. Mr M.B. Line and Mr R.G. Bradshaw read the draft version and made many suggestions for improvement.

1 INTRODUCTION

1.1 Statement of aims

The aim of this part of the DISISS project is the application of statistical techniques to citation data in order to group journal titles in the social sciences. Various statistical techniques exist - some of them with a fairly long history - for grouping items according to observable attributes. Details of these techniques and the selection of one for use with DISISS data are discussed in 4.

1.2 Data to be used

Data from two computer files is being used for the clustering work: (a) citations taken from ISI Science Citation Index (SCI) for one-quarter of 1971; and (b)citations gathered by hand from social science journals.

1.3 Schedule of work

The cluster program has been developed using data collected during the pilot citation study in 1971. This data was collected from 17 source journals (see Appendix A) and partially analysed and presented in Working Paper no. 5. Further details of the data used for the preliminary clustering work are given in Appendix B and described in section 5.

For the main work, data is being extracted from tapes of Science Citation Index and transformed for use with the cluster program. This data from SCI covers several social sciences journals, with a predominance of psychology. It will be merged with the data collected by hand for the main DISISS citation file and the full clustering runs will be carried out on the combined file.

1.4 Reasons for using citation data

Since few subject boundaries in the social sciences are clear cut, and because the field is changing all the time, any secondary service which attempts to cover less than the entire field is faced with the problem of defining its scope. Traditional methods of classification have two weaknesses, their inflexibility and their reliance on human judgement. In particular, subject areas which are involved in interdisciplinary work are likely to be separated in any sort of classification scheme which is based on a subject hierarchy. This study is an attempt to devise a flexible method of dividing up the social sciences objectively, based on the behaviour of people working in the field. Ideally we would like to separate the literature into clusters such that all items of relevance to any individual would be in a single cluster, but, since this is impossible, the aim is to divide the literature into clusters which come as close to this goal as possible.

Citations, although far from ideal, are suitable data for the following reasons.

(i) The data is easy and relatively cheap to obtain.

(ii) Citation data is available across a wide range of subjects, including all important social sciences.

(iii) There is a positive relationship between use and citation.[1]

For this study we have restricted the data to citations from journals to journals. This is purely for convenience and has some disadvantages. The following are the main disadvantages of using citations for this work.

(i) Value, use and frequency of citation, although related are not identical and one does not automatically describe the others.

(ii) Where subjects have developed on parallel paths there will be few cross citations even where relevant.

(iii) Much work in the social sciences is not published in academic journals, but in monographs, reports or journals which do not contain citations.

(iv) Some subject areas have different patterns of citation from others.

---

[1] The precise strength and nature of this relationship remain to be studied.

Above all the reasons for using journals citations is that it is the only type of data for which large amounts of hard, objective data can be gathered at all cheaply.

2 PREVIOUS WORK

Section 4.2.3 reviews those parts of the literature on clustering techniques which are relevant to this study, both from a statistical point of view and with reference to DISISS.

Clustering, grouping, and clumping techniques have been applied to three types of bibliographical data: (a) citations; (b) index terms, which may be derived from titles, abstracts or very rarely from a text, or maybe assigned to a document; and (c) terms in user requests. Clustering work involving index terms usually has the objective of automatic classification of documents. This is also the objective of some work on clustering documents by citation patterns, but clusters of journal titles are of most interest in dividing the journal literature into groups according to the judgements of researchers themselves (as indicated by their citation practices). Work on clustering user terms usually aims to match queries with groups of documents identified in retrieval, which may themselves have been identified by grouping of terms in documents.

Much more work has been done on automatic classification of documents according to index terms than on grouping documents by citation patterns. Important reviews are those by Stevens (1965), Salton (1968), and Stevens, Giuliano and Heilprin (1965). These reviews mention clustering by citations, but deal mainly with grouping of index terms. Important work in grouping index terms is by Salton and Borko (1965), Borko and Bernick (1963), (1964), Dale and Dale (1965), Doyle (1962), (1964), Gotlieb and Kumar (1968), Sparck Jones (1971), and Stiles (1961). Three studies using index terms have been reported by Williams (1966), Wolff-Terroine and Rimbert (1968), and Augustson and Minker (1970). Worona (1969) uses the terms in user requests as well as documents, and develops a query clustering procedure. Interest in classification and grouping of user requests is more recent, and neither this work, nor the work on the grouping of documents by index terms will be discussed further in this paper.

There is relatively little work to discuss about the application of clustering techniques to citation data. Little attention has been given to the statistical techniques, the requirements which citation data makes, the conditions which apply to the statistical techniques, and the extent to which citation data can meet them. Failure to meet basic requirements gives rise to problems in the interpretation of the results.

Interest in the use of clustering techniques for the organization of bibliographical material has largely arisen in the last five years. Five studies are of particular interest.

Xhignesse and Osgood (1967) applied an interpoint distance procedure to a citation data matrix consisting of psychology journals; the method produced clusters of interacting journals. The method was tried out on a symmetrical universe of source journals (the 21 source journals also being the only cited journals allowed) and the results obtained appear to be useful.

Price and Schiminovich (1968) used a clustering procedure with a bibliographic coupling measure on a collection of 240 theoretical high energy physics papers. This study was intended as a first step towards producing an entirely automatic classification scheme. Later work by Schiminovich (1971) was designed to overcome some of the failings of the simple approach used in 1968, which appears to have been discarded. They developed a "bibliographic pattern discovery algorithm" using more information on the bibliographic links between documents than is contained in the simple bibliographic coupling measure. This algorithm forms the basis for a method of generating groups of papers and a classification system for documents in the subject area. When applied to a collection of about 30,000 physics documents, the groups of documents generated corresponded to recognizable topics even when spread over several conventional classification categories. The generated classification system compared favourably with one used by a journal for indexing its articles.

Carpenter and Narin (1972) have used a cluster analysis procedure in which 288 highly cited journals in physics, chemistry,

and molecular biology were grouped into clusters of related journals. The clusters could be identified by national, subject and subdisciplinary divisions.

Narin, Berlt and Carpenter (1972) have also investigated hierarchies of journals, using the two journals most cited by each journal in the study, to divide chemistry journals into groups. This first stage can be done by hand, and results in groups containing a manageable number of journals for further clustering. In other words, journals are divided roughly into subject, discipline, or interest groups and cluster analysis can then be undertaken on each group. This is much easier than performing the clustering initially on a large heterogeneous group of journals, partly because of the computational problem of dealing with more than 100 source journal titles in a cluster, and partly because of the difficulty of interpreting clusters derived from a large, heterogeneous set of titles, covering many subject fields. This method is not applicable to DISISS, because one of the objectives is to investigate the value of cluster analysis, in dividing the subject matter of the social sciences, where there is less agreement about discipline boundaries than in the physical and biological sciences.

3 USE OF CLUSTER TECHNIQUES IN BIBLIOGRAPHY

3.1 Implications of clustering for the design of information systems

The main objective of applying cluster analysis to bibliographic data, in the form of journal titles, authors or papers, is to develop and identify groupings which can be used to structure bibliographic files. DISISS deals with cited journal titles data. The data obtained from clusters of journal title citations can be applied in the following ways

(i) It can assist in planning patterns of journal coverage for secondary services. The clustering process defines groups of journals by classifying them into related subject groups. The rationale for applying the results of cluster analysis to the design of secondary services is based on the fact that the citations used in clustering are user generated. The designer of a secondary services thus has a measure by which he can attempt to match the supply of information to the use of, and possibly the need for, information. This applies to existing services as well as providing a basis for new services.

Subjective methods (e.g. consensus of users, editors, experts, local availability of material) used to decide journal coverage contain no element whereby the relative importance of journals can be measured. On the other hand, citation data is objective and has the merit that it indicates what users read, later write up and cite.

Certain assumptions are necessary to justify the use of citation data; citation practices do not necessarily neatly match use and probably less so in some of the social sciences than in science; nor by any means do users meet all their bibliographical requirements from secondary services. The representativeness of the citation data base on which cluster analysis is carried out also needs to be considered; does it provide a valid

random sample of the literature, what are the effects of studying different time periods, etc? Until more is known of the distribution of citations, it is impossible to estimate the effect of different sampling procedures, but some simple checks can be carried out.

(ii) Journal title clusters provide data on the structure of the primary literature in a discipline. From this general picture:

(a) new fields can be identified, either by comparison with earlier results or because unexpected patterns are displayed;

(b) the need for new or modified services can be seen and appropriate action taken;

(c) hypotheses can be developed which may lead to new lines of enquiry and development of services.

(iii) Cluster patterns can give evidence of the validity of thesauri, and classification and indexing schemes.

(iv) Cluster data is not only relevant to the operation and design of secondary services; the data might indicate where the primary literature could be rationalized, or even expanded, although some data on the material contained by the journals in clusters would be required.

(v) Clustering allows various descriptive studies of the social science literature to be carried out within a firm framework of subject groups. An example of such studies would be the growth and obsolescence rates of subject groups. Other studies of interest are language, and country of origin of journals in each subject group. These studies may be carried out as comparisons between subject groups and within each subject group.

4 TECHNIQUES OF CLUSTERING

4.1 Clustering terms

A clustering approach or technique is a method of grouping data points without using a preset classification. A clustering algorithm is a description of such a method which is sufficiently well-defined to be programmed or executed manually without further definition. A hierarchical clustering or classification is such that at the lowest level, each data point is a separate cluster, and at higher levels, each cluster is formed by merging complete clusters from a lower level. At the highest level the whole data set forms a single cluster. Such a clustering can be shown in diagrammatic tree form as a dendrogram (e.g. the figures in Section 5.4).

A multivariate data set is one in which each point is defined by values on a prescribed set of variables. A similarity matrix for a data set is a matrix of positive elements, S_{ij} being the similarity between point i and point j. Usually it is assumed to be symmetric ($S_{ij} = S_{ji}$) with similarities in the range [0,1]. In this case $S_{ij} = 1$ implies that i and j are identical within the terms of the data and the matrix has only n(n-1)/2 significant entries since clearly $S_{ii} = 1$. A dissimiliarity matrix is similar but $S_{ii} = 0$ and the range of the dissimilarities is seldom restricted.

4.2 Clustering techniques

Clustering analysis differs from other statistical techniques such as regression or the analysis of variance in that it is not a well-defined technique. The term is applied to a number of widely different approaches, whose only common feature is the objective of identifying groups of points within a given set of data points. This is as far as similarity goes. Some approaches try to detect a single isolated group at a time, which is distinct from the rest of the set. Others try to divide the set into groups to optimize some overall

criterion. Yet other methods impose a hierarchical structure on the data. The definition of a cluster, the criterion for optimization or the rules for forming a hierarchy can be varied to stress different qualities required of the clusters: isolation, connectedness, compactness. The purpose for which the clusters are to be used, and hence the requisite qualities and structure of the clusters are clearly important when selecting a clustering technique.

The form of the data required by an approach can be either a multivariate array (where each point is described by its values for a fixed set of variables) or a similarity (or dissimilarity) matrix (where a similarity is specified between each pair of points). It is always possible to convert a multivariate array into a similarity matrix by defining a suitable distance function (e.g. Euclidean) of the variables for a pair of points, but this process has two disadvantages. The important disadvantage is that where the number of variables is small relative to the number of points the similarity matrix is much larger than the equivalent multivariate array.* In addition, if it is conceptually reasonable to picture the points in n-dimensional space (particularly if the variables are comparable and reasonably independent) some useful information may be lost in the conversion. In particular the facility is lost for directly identifying a cluster by a hypothetical point (e.g. the centroid of the points in that cluster), which can be a computational or a conceptual aid. The advantages of the similarity matrix appear when there is no convenient fixed set of variables (e.g. when using co-occurences of keywords in their titles to measure the relationship of two articles) or when the variables are not comparable (e.g. height, weight and age). In the latter case the distance function, which will be required for a technique using the multivariate array, may be so complex and time comsuming to evaluate that it is more efficient to determine the similarity matrix once and for all

---

*For N points defined by n variables the size of the multivariate array is N×n. The size of the similarity matrix obtained from this is $\left(\frac{N(N-1)}{2}\right)$. Hence if $n < \frac{N-1}{2}$, the similarity matrix is larger than the multivariate array.

rather than continually evaluate distances within the main algorithm. Clearly, the form and amount of the data which is available, or which can be collected conveniently, must be also considered when selecting a clustering technique.

Another reason for taking the amount of data to be clustered into account is that clustering techniques vary considerably in the time taken to cluster a given number of points. Any approach which requires complex calculation for its basic step is likely to be unusable for large amounts of data. For instance, an approach using a complex overall criterion for optimizing the division of a set into clusters will not be efficient if it has to be wholly or largely recalculated whenever a decision has to be made whether to move a point from one cluster to another. In particular, if the multivariate array or similarity matrix is too large to fit into the computer core store and some form of paging is required, it is essential to use a technique which does not require the whole matrix for every basic step.

4.2.1 The SCICON algorithm

The clustering method chosen is based on a non-hierarchical method developed at SCICON by E.M.L. Beale and M.G. Kendall. This method, referred to as the SCICON method where confusion might arise, assumes that the data is in the form of observations, or data points (in our case cited journals) consisting of measurements on each of a fixed set of variables (in our case citing, or source, journals). If the number of variables is n, these observations are considered as points in an n-dimensional Euclidean space.* At any stage, a point is allocated to one and only one cluster. The

---

* In a Euclidean space, if the ith point X_i is represented as $(x_{i1}, x_{i2}, \ldots, x_{in})$ where x_{ik} is the measurement of observation i on the kth variable, the distance between X_i and X_j is defined as

$$\left(\sum_{k=1}^{N} (x_{ik} - x_{jk})^2 \right)^{\frac{1}{2}};$$

i.e. Pythagoras's theorem extended to n dimensions.

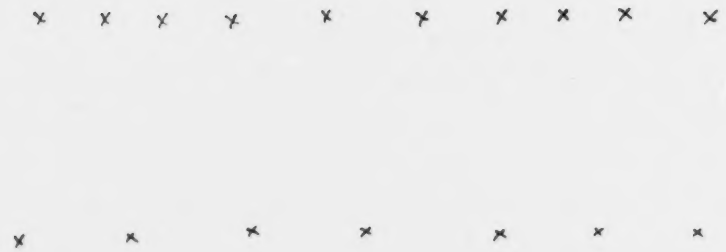
criterion used to decide which of two different divisions of the points into a number of clusters is the root mean square deviation from the centres of gravity of the clusters (see Appendix D). This is a measure of the average distance of points from the centres of gravity (assuming equal weights for all observations) of the clusters to which they have been allocated. The better allocation will have a lower value for the criterion.

The basic idea is to divide the points into clusters which are so chosen that for a given number of clusters, the "average" distance from the points to the centre of the cluster to which they are allocated is minimized. If this criterion is evaluated for divisions into different number of clusters, the optimal division will be better for the larger number of clusters. The criterion is therefore not applicable to comparisons of allocations to different number of clusters. The procedure used is that, given an allocation to m clusters, each point is examined in turn to determine whether moving it to any other cluster will improve the criterion. If so, it is moved and the next point (in order of submission to the program) is then considered for reallocation. When a full pass through the points performs no reallocations the current allocation is output as the best division into m clusters that has been found. This may not be the optimal allocation of the points to m clusters that could be found by total enumeration, but this would be impossibly time-consuming for more than a trivial amount of data. The factors which restrict the optimality are the initial allocation, the order in which the data is submitted and the consideration only of single points for reallocation rather than allowing groups of points to be reallocated simultaneously. With the number of source and cited journals that we have it is not feasible to alter the last factor because of limitations of computer storage and run time. The initial allocation to clusters is described below.

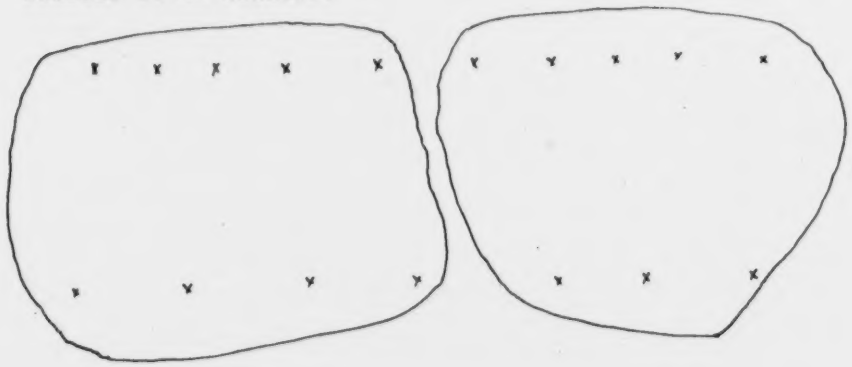
In any run of the program, the method is non-hierarchical; the algorithm produces a sequence of divisions into a decreasing number of clusters. The initial allocation for the m - 1 cluster stage is obtained by merging the two closest clusters in the best allocation obtained for m clusters. This procedure means that computation for later stages is often small and that, although not strictly hierarchical, the sequence of cluster allocations obtained often resembles a hierarchical structure.

The allocation of points to clusters at the beginning of the first stage is made as follows. A sequence of cluster centres is formed by first selecting the data point furthest from the centre of gravity of the whole data set as a cluster centre, and subsequently choosing as the next cluster centre the point which maximizes the minimum distance from the point to any of the previous cluster centres. Each point is then allocated to its nearest cluster centre. This procedure has been found by experience to yield quite a good initial allocation, while simple to program and quick to execute. It is advisable to use a higher value for the maximum number of clusters to be considered than is strictly required, to allow a running-in period of several stages. This running-in period should usually counteract any effect of the order in which the data is submitted and the technique of choosing the first set of cluster centres.

It will be seen that this method is essentially a pragmatic method which experience has shown gives consistently good results with a variety of real life data. It does not require data to follow any particular statistical law, and freak sets of data can easily be imagined which it would not handle. For instance, in two dimensions, the following data



appears to fall into two groups, but the algorithm might cluster it as follows.



4.2.2 Reasons for selection of the SCICON algorithm

In DISISS we wish to cluster a large number of social science journals, perhaps two or three thousand. A hierarchical structure, while an obvious advantage in designing automatic retrieval systems for documents, would be an unnecessary restriction in the design of secondary services. However, if a method, which does not impose such a structure, were to reveal one, this would be valuable information. Social scientists use and cite much material outside their immediate discipline. It seems therefore more reasonable to attempt to divide the set of journals into clusters which optimize some overall criterion than to identify single groups which are distinct from the rest of the set. The latter approach requires a definition of a cluster which is liable to be either so stringent that only a few small groups of journals are sufficiently isolated to satisfy the definition, or so lax that the number of clusters produced by the algorithm is so large that the analysis required after clustering would be enormous.

The data was limited by what could be collected conveniently and at a reasonable cost. Although some data was available from Science Citation Index the social science source journals covered by this service were mainly in the field of psychology and, to cover the rest of the social science disciplines, it was essential to collect most of the data independently. It was impossible to decide which journals were to be clustered and to collect citations to all of them, from all of them, for three main reasons.

(i) Part of the aim of DISISS is to identify the important social science journals, and since the citations collected are partly for use in this identification, a circular situation arises.

(ii) Copies of foreign and specialist journals are often difficult to track down and it is of interest to cluster these.

(iii) The cost of collecting sufficient citations from all the journals to be clustered is prohibitive.

The alternative procedure is to select a few source journals and collect a large number of citations from each, and then clustering all the journals cited by these (in practice it is reasonable to consider only those journals cited more than a certain number of times). The choice of source journals is dealt with in section 4.3. Regarding each source journal as a variable, we have a multivariate array with each cited journal being a point defined by the number of citations from each source journal. Obviously the number of source journals will be small compared with the number of cited journals, and the variables, all being numbers of citations from a source journal, are comparable if not independent.

From these considerations it is reasonable to choose a method which uses a multivariate array to divide a set of points into a number of clusters. The SCICON method satisfies these requirements. It has several other advantages, one being its availability in the form of a program for the ICL 1900 series computers at the University of Sussex and at the Open University. Although, in common with most algorithms which use an overall clustering criterion, it does not guarantee an optimal solution, there is a facility for clustering over a range of decreasing numbers of clusters and experience with a variety of real data has shown that although the clustering obtained for the first two or three numbers in the range may be definitely sub-optimal, after this initial run-in near-optimal clusterings are obtained.

Because the change in the criterion (mean square deviation of points from the centre of gravity of their cluster) when a single point is transferred from one cluster to another is simple to calculate using only the position of the point itself and the two relevant centres of gravity (see Appendix D), the algorithm can cope efficiently with quite large numbers of points, even if paging of the main data matrix is required.

Using the facility for clustering over a range of number of clusters it is possible to detect a distinct hierarchical structure if one is present by checking whether a reduction in the number of clusters results in the amalgamation of whole clusters, or in the redistribution between several clusters of points which had been in one cluster. Clusters which remain unchanged over several reductions in the number of clusters are clearly fairly distinct and isolated. Useful subsidiary information which is available for each clustering is the ratio of the distance of each point from its second nearest cluster centre to the distance of each point from its own cluster centre. This is a measure of the adhesion of the point to its cluster. If the ratio is not far from unity the point may fit nearly as well into the second cluster. This ratio may be incorporated in an extended method allowing overlapping clusters. Other information which might be useful conceptually is the location and hence relative location of the cluster centres, and the mean distance of points in a cluster from its centre, which is a measure of the density or compactness of a cluster.

4.2.3 Comparison with other approaches and algorithms

To compare the SCICON method in detail with all clustering approaches that have been suggested would be an enormous task. This section covers some of the most widely known and some which have been applied to the specific task of grouping journals. It also attempts to highlight some of the main differences in approach and the reasons for these differences.

Over the last twenty-five years considerable effort has been exerted in developing automatic techniques for dividing a data set into natural or homogeneous groups without previously specifying the groups in any way. This effort has been spread over a number of disciplines: biology (where it is usually known as numerical taxonomy); psychology (e.g. for grouping subjects according to experimental results); linguistics (e.g. for classification of

phonemes); information retrieval (e.g. for classification of documents in a collection); and marketing (e.g. to identify groups of competitive products and gaps in the market which might be filled with a new product). The different contexts impose different constraints on the form of clustering required, and the form of data available. Both these factors impose requirements on clustering techniques and the algorithms which implement them. It is unlikely that any single approach could satisfy the full range of requirements made in all circumstances. The following may highlight some of the differences:

(1) Is a hierarchical structure necessary (e.g. for definition of a tree structure, for efficient automatic information retrieval, or for biological taxonomy)?

(2) Is the clustering required to group the whole data set? Can outliers be ignored? Can overlapping clusters be allowed? Is it required merely to identify individual clusters which are distinct from the rest of the data?

(3) In what form is the data?
   (i) similarities between data points (e.g. co-occurrence of key-words, numbers of confusions of two phonemes)
   (ii) measurements on comparable variables (e.g. counts of citations from journals to each other)
   (iii) measurements on a set of variables which are not directly comparable (e.g. height, weight and age of individuals).

(4) How much data is there (20, 100 or 1000 points)?

The fourth of these questions probably implies a further one. Is the exact clustering of individual points important, or

merely the statistical properties of the overall grouping? The position of an individual plant in a small plant taxonomy is different from the position of an article in a large document collection, where the statistical retrieval performance is the important factor.

The SCICON method is clearly unsuitable if a strictly hierarchical structure is required. It is not obvious whether a hierarchical method is appropriate for classifying social science journals. To impose a hierarchical structure on data which has no inherent hierarchy is undesirable, but if there are indications of such a structure it is certainly of value to compare the results of hierarchical methods with those of the SCICON algorithm (see sections 5.4 and 5.5). The most common family of hierarchical methods can be described as follows.

(a) Input the $\frac{n(n-1)}{2}$ similarities between the n points to be clustered.
(b) Consider each point as a separate cluster.
(c) Choose the two closest clusters p and q and merge them into a new cluster k which replaces p and q.
(d) Compute the distance of cluster k from any cluster s, as a function of the distance of p and q from s.
(e) Return to step (c).

The process clearly stops when the last two clusters are combined. Associated with each merger is the distance between p and q when they are merged. The distinguishing factor between members of this family is the distance function computed in (d). Cunningham and Ogilvie (1972) list 7 variants and Anderberg (1971) discusses 5 of these and adds another. They are listed in Table 4.2.1 below.

Table 4.2.1

Distance functions suggested for hierarchical clustering

(1) Single-linkage (nearest neighbour)
(2) Complete-linkage (furthest neighbour)
(3) Average linkage between merged groups (group average)
(4) Average linkage within new group
(5) Centroid
(6) Median
(7) Ward's error sum of squares
(8) Simple average.

Variant (7) in Table 4.2.1 is closely related to the SCICON criterion and Anderberg (1971) has experience which suggests that its results when evaluated using the SCICON criterion are good. Compared with the SCICON algorithm its disadvantage is that it requires to start with n clusters where n is the number of data points, and is therefore very time-consuming for large data sets. Any alteration to the implementation to counteract this would leave something very similar to the SCICON algorithm with the disadvantage of not allowing redistribution between clusters to improve the criterion value.

Cunningham and Ogilvie (1972) suggest that on several sets of artificially structured data (3) was consistently good at revealing the structure,with (2) and (8) producing similar results (Anderberg (1971) suggests that (4) performs similarly). They found that (5) and (6) tended to distort input data that was strictly ultrametric and thus hierarchical[1] but performed reasonably on data describing not very distinct clusters in the Euclidean plane.

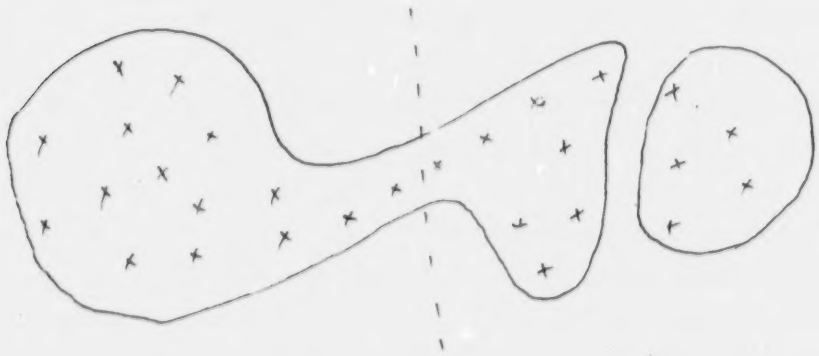
The function related to the SCICON criterion (7) performed very well on all types of data except one. The criterion favours clusters which do not differ greatly in size and (7) did not reveal

[1] Johnson (1967) develops a correspondence between a hierarchy and similarity matrix satisfying the ultrametric inequality

$$d(x,y) < \max[d(x,z), d(z,y)]$$

the structure of hierarchical data which represented such clusters. This tendency is not a disadvantage when the aim is to use clusters in the design of secondary services. One large cluster together with a few very small ones would be difficult to interpret and use rationally. However it implies that the criterion is not suitable for use when such a structure is suspected and the hierarchical position of the individual items is important.

Single-linkage (1) is probably the most widely known and used approach, largely because it is very simple to apply efficiently to small or large sets of data. Its disadvantage, which is often considerable when clusters are not very distinct, is a tendency towards 'chaining'. Chaining is best illustrated diagramatically.



Although the intuitive way of splitting the above data into two clusters might be to divide it along the dotted line, single-linkage will tend to form the left-hand cluster and then follow the central chain of points incorporating one at a time, thus producing the two circled clusters. single-linkage can therefore easily produce long straggly clusters in which extreme points are very dissimilar. For some, but not all, applications this is a disadvantage. Methods (1) and (2) have the conceptual advantage that the resulting clusters strictly maximize intuitive properties of connectedness and compactness respectively. At a given stage of single-linkage clustering, identified by a distance r, all points that can be joined by a chain of points

less than r apart are in the same cluster. Any stage of complete-linkage clustering the maximum distance between any two points in the same cluster is s. As the algorithms work up the hierarchy, r and s increase steadily. Complete-linkage (2) therefore concentrates on compactness and is probably more suited to applications where the overall picture is more important than the positions of each individual. This is borne out by the results in section 5.4. These two methods in one sense represent extremes between which results from the other five may be expected to fall.

One feature of single- and complete-linkage which is often stressed is their invariance under all monotone transformations of the similarity matrix. A monotone transformation is one which preserves the ranked order of the elements of the matrix. Johnson (1967) among others has suggested that psychologists only have sufficient confidence in their similarity data to say that if $d(a, b) > d(p, q)$ then a and b are more similar than p and q. They may not be prepared to concede any significance to the size of the difference between the similarities. For large sets of data with similarities restricted to a finite range (as is often the case) this lack of confidence loses its significance. Hubert (1972) suggested that in fact psychologists would be prepared to concede more confidence in their data, to the extent of the rank-order of the first differences of their similarities. Transformations which preserve this more constrained rank-order are called hypermonotone. He suggests a simple extension from single- and complete-linkage which is invariant under hypermonotone transformations, though not under a general monotone transformation, defining the distance between two clusters as the sum of the minimum distance between a point in one cluster and a point in the other and the maximum distance between such pairs of points. His experience shows as expected that this produces results intermediate between single- and complete-linkage. Unfortunately it is more difficult to implement efficiently than either.

One point worth making at this stage is the distinction between a method and its algorithmic implementation (stressed by Jardine, 1970). Although the seven methods listed in Table 4.3.1 can be described as variants of a simple process, often the most efficient algorithm for

obtaining the results is by a completely different sequence of steps. Efficiency is affected by two restrictions, time and space, and which is the most efficient algorithm (measured by total 'cost') may depend on the size of the data and on the particular computer[1].

R.F. Ling (1972) has suggested a method which he regards as a generalization of the single-linkage method. As with single- and complete-linkage his definition of a cluster has a conceptual basis though it is somewhat more complex. He defines a (k, r)-cluster S roughly as follows.

(i) Each point in S is a distance less than r from at least k other points in S [(k,r)-bonded].

(ii) Any two points in S can be connected by a chain of points in S in which each link distance is less than r [(r)-connected].

If k is chosen as 1, this method yields the single-linkage method. Complete-linkage requires clusters of k elements to be (k-1,r) bonded for some r.

For any fixed k, the set of (k,r)-clusters in a set of data forms a hierarchy. However the hierarchy differs from those discussed earlier in that a merger may be of several clusters rather than just two. To take full advantage of the promising features of this method it would be necessary to compare the hierarchies obtained for different values of k. This would of course be time-consuming, especially as run times increase exponentially with k. A program has been obtained from Ling and it would be interesting to investigate its results on a small data set.

Before leaving the topic of hierarchical methods it is worth pointing out that some of the least desirable features for hierarchical methods listed by Jardine and Sibson (1968) apply only to applications where the exact position of individuals is the important feature. Since the single-linkage method is the only one found to possess these features this is reassuring.

---

[1] Sibson (1973) has recently published an efficient, single-linkage algorithm for very large data sets.

Non-hierarchical methods are considerably less easy to classify and compare than the hierarchical ones. They have often been designed with particular applications in mind and concentrate on producing good results for a particular type of data. Several methods have been designed principally for binary (dichotomous) data of presence or absence variables and though theoretically applicable to transformed multi-state or continuous data the results may be harder to interpret in the light of the required transformations.

Harrison (1967) describes a program developed in ICI to detect clusters of biologically active compounds using data on chemical structure. It uses binary data and a probabilistic measure of clusters significance to detect patches of greater than random density of active compounds. It does not attempt to partition the whole data set into clusters and although an adaptation of the ideas might possibly be appropriate to DISISS the task of programming and determining run parameters by experimentation would have been considerably greater than for the SCICON method which was already available as a program compatible with the DISISS data.

The broad concept behind most non-hierarchical methods is to start with an initial partition of the data, and then move points between clusters to obtain a better partition. Methods differ as to what constitutes a better partition and what techniques are used for improvement. Where an explicit criterion, such as the SCICON criterion[1], is used, a hill-climbing algorithm is usually employed, possibly with modifications. Roughly, such an algorithm allows only uphill steps, steps which positively improve the criterion. Such algorithms are always liable to find local rather than global optima. It might be better in the long term to go down into a valley to reach a higher peak. The SCICON program uses a hill-climbing algorithm at each cluster level and can only guarantee a local optimum at any level. However experience has shown that near-optimal results are usually obtained by using a run-in period of about 5 cluster levels. Although absolute optimality could not be checked for the results in section 5.3, consideration of the criterion values after various

[1]This criterion is sometimes referred to as trace W, the trace of the pooled within groups scatter matrix, as by Friedman and Rubin (1967).

run-in periods using the pilot study data tends to corroborate this claim. Other explicit criteria for multivariate data have been discussed by Anderberg (1971), Friedman and Rubin (1967), and by Scott and Symons (1971). Most of these have the disadvantage of requiring complex and time-consuming calculations in testing whether moving a point will improve the criterion or not. Very little published work has been done using them. Rubin (1967) introduced another criterion using a concept of 'average object stability', where the stability of an object in a cluster is a measure of the similarities of a point to the other points in its cluster compared with its similarities to points in another cluster. This criterion is based on a similarity matrix rather than directly on multivariate data, and is comparable for division into different numbers of clusters. This last feature means that Rubin's modified hill-climbing algorithm finds an optimum number of clusters as well as an optimum partition into a particular number of clusters. However his criterion also involves a variable parameter and runs for several values would be necessary for useful results[1]. He suggests a number of tactics to apply in an attempt to improve on local optima but remarks that they are only occasionally helpful. This method,which would be very complex to program, has been little studied. This neglect may be due to its complexity rather than to any obvious defect.

Several people have suggested algorithms which while not optimizing an explicit criterion implicitly use one very similar to the SCICON criterion. Anderberg (1971) discusses several of those which, given the centroids of a partition, allocate points to their nearest centroid, recompute the centroids and recycle. The algorithmic details vary and different methods of improving on local optima or optimizing the number of clusters are suggested but none of them shows any obvious advantage over the SCICON algorithm.

There are a few published examples of attempts to group journals. These have tended to use rather ad hoc methods which might be difficult to apply to large amounts of data. Brief descriptions of three attempts follow.

---

[1] A very recent paper by Gitman (1972) suggests a similar approach involving no run-time parameters.

Xhignesse and Osgood (1967) used citations within a group of 21 psychology journals to obtain a similarity matrix (the exact calculation used is not described). A method developed by Shepard (1962) was applied to this matrix to represent the journals in multidimensional Euclidean space in such a way that the rank-order of the similarity elements was preserved. This is not a clustering technique; it is a method for obtaining a Euclidean multivariate representation of data which naturally produces a similarity matrix. After applying this process Xhignesse and Osgood used an arbitrary distance to define overlapping clusters of journals. The exact procedure is not stated but it is probably based on defining clusters as groups of journals lying within spheres of a certain diameter. It is most unlikely that this method, which is rather arbitrary and ill-defined (at least in the published paper), would be applicable to large amounts of data.

Parker, Paisley and Garrett (1967) found clusters among 68 journals in the field of communications. Unfortunately they did not describe the clustering method in detail and omitted the relevant references from their bibliography. Citations were taken from 17 source journals and co-occurrence of citations to pairs of journals from the articles in these sources was used as a measure of similarity between the journals. This measure is almost an inverse of bibliographic coupling as introduced by Kessler (1963a) and discussed later in this section. A disadvantage of this measure is that, without any adjustment for the level of citation to each journal, pairs of highly cited journals are almost certain to have high similarity. The clustering approach appears to identify a tight cluster of highly related journals, and then discard the journals in that cluster from consideration. It is difficult to see what application of these clusters is possible especially if journals from more than one field are considered.

Carpenter and Narin (1972) have followed their earlier study (Narin, Carpenter and Berlt, 1972) of the structure of the journal populations in scientific disciplines through relative citation levels between journals, by clustering journals according to citation data.

Their clustering approach involves a similarity matrix and a simple hill-climbing method to optimize an overall criterion. They do not specify whether the choice of the number of clusters is internal to the algorithm or chosen beforehand. They have been unable to decide on a best similarity measure or clustering criterion and, making the reasonable assumption that clusters which are detected by several different methods are definitely clusters, they combine the results using several similarity measures and clustering criteria. This is done by producing a new similarity matrix in which the number of times journal x appears in the same cluster as journal y is the similarity between the two journals. A single linkage clustering is performed on this matrix to obtain the final clusters. This combined procedure is an admirable approach; although it would be difficult to implement on the whole of the DISISS data, it has many advantages for smaller samples.

The basic data used by Carpenter and Narin is cross-citation between all the journals under consideration, making use of Science Citation Index tapes. It is not possible to reproduce their analysis using social science data because the projected Social Sciences Citation Index is not yet available.[1] It might be of interest to use their methods of analysis for those source journals in the ranked and random lists for the main DISISS citation study. However it is possible that 75 journals spread over the whole of the social sciences may be too small a set to achieve results that are not self-evident anyway.

Finally a few comments can be made on some attempts to group documents, usually journal articles, using citations made by these documents.

Preparata and Chien (1967) suggest an algorithm for solving the problem of arranging documents (or document descriptions) in sequence to minimize, in some sense, the distance between similar articles. This could be used to minimize access time for documents retrieved in response to a query, where the collection is held on magnetic backing store of a type that requires movement of the reading heads over the file.

---

[1] The Social Science Citation Index is now available.

They consider a similarity matrix in which the similarity between two documents is "1" if either document cites the other, and "0" otherwise. The algorithm uses a hill-climbing approach to minimize the total distance between all pairs of documents with similarity 1. Although the details are very different the approach is related to a restriction to one dimension of the scaling method of Shepard (1962) used by Xhignesse and Osgood (1967) and referred to earlier. Despite the title of the report[1] by Preparata and Chien this is not clustering in the usual sense of the word. There is no other obvious application of the solution to their problem, and in a computer system it would probably be possible to reduce access time further by an arrangement of the collection specific to the particular configuration.

Kessler (1963a) introduced the concept of bibliographic coupling between articles and used it to define two criteria for grouping papers. The bibliographic coupling between two papers is the number of items cited by both papers. High coupling tends to indicate a strong relationship between articles because they refer to common work. As it stands, this measure is not suitable as a similarity measure for the clustering techniques described in the earlier part of this section; papers which cite heavily are more likely to be highly coupled with other papers, and the coupling of a paper with any other is limited by the number of citations it gives. No suitable modification of the concept to counteract this effect is immediately obvious, but the approach might be interesting to investgate. Kessler uses the concept to generate two types of group. The first type of group is generated from a triggering paper and consists of all papers in the set under consideration which have at least n citations in common with the triggering paper. The second type of group is one in which each paper has at least n citations in common with every other paper in the group. This second condition is a stringent one, somewhat related to that suggested by Van Rijsbergen (1970), and seems to have been largely ignored by Kessler in later papers (1963b, 1965). Using the first criterion any paper can be used as a triggering paper to generate a group of papers from the set under consideration.

---

[1] F.P. Peparata and R.T. Chien (1967) Report R-349
Coordinated Science Laboratory, University of Illinois, May 1967.
"On clustering techniques of citation graphs".

Not all papers will necessarily trigger off a group at all and in that case they will not be in any generated group since they cannot share n citations with any other paper in the set. However the groups which are generated will certainly overlap, without overlapping groups necessarily being identical. This form of grouping or clustering is less suitable for DISISS than for designing a classification system to be used explicitly or implicitly within an information retrieval system.

Schiminovich (1971) found the simple idea of bibliographic coupling expressed as a numerical value unsatisfactory for developing a fully automatic classification system. He used both the numerical value and the information of which citations contributed to this value to generate a sequence of groups of papers which should converge to form a bibliography which could define a class in a classification. He used Kessler's idea of a triggering paper but realized that the trigger need not actually be a paper with its list of citations. Any list of citations could serve as a trigger, e.g. a user-provided bibliography, or a group of papers produced by an earlier stage of his automatic process. The pattern discovery algorithm which forms the basic step of his method could be adapted to cluster documents using index terms or word content and could probably be usefully applied to completely different problems in different fields.

4.3 Data for clustering

DISISS is using citation data for a variety of analyses of the citation practices of social scientists, one of which is to cluster social science journals. In section 4.2.2 some practical reasons were given for collecting a large number of citations from a few source journals as clustering data. These reasons also apply to the data for other analyses. As remarked in section 5.3, for clustering it is desirable to adopt one of two strategies:

i) collect equal numbers of citations to journals from each source journal;

ii) collect all citations (or every nth) from all articles (or every mth) are a fixed time period.

For some of the other analyses citations to all forms of material are required, in the proportions in which they occur, and in such a way that numbers of citations from different source journals can be combined. In this case two reasonable strategies would be:

i) collect equal numbers of citations from each source journal, and count total number of citations from each source journal;

ii) collect all citations (or every nth) from all articles (or every mth) over a fixed time period.

The second strategy is identical in both cases, whereas the first strategies would involve identifying two different but overlapping sets of citations to journals. The second strategy is also easier to apply in practice since collectors can work independently and without keeping an exact count of citations as they work. It was therefore decided to collect every citation from every article in 1970 for source journals from which citations were collected by hand. Citation collection is time-consuming and since magnetic tapes of the citations in Science Citation Index for one quarter of 1971 were available it was decided to make use of this data source where possible. For source journals on these tapes all citations for one quarter of 1971 are available. It is necessary to use a weighting factor when combining citations from the two sources.

The choice of source journals could have considerable effect on any analyses of citation data. Some thought has been given to this problem (see Working Paper 5) and it was decided to select source journals in three ways.

i) Select the fifty journals most cited in the pilot study.

ii) Select a random sample of fifty journals from CLOSSS.[1]

iii) Select key journals in subject or language categories not covered by i) e.g. public administration, geography, Russian.

The complete list of journals selected as sources is in Appendix C.

---

[1] CLOSSS = Check List of Social Science Serials. This list has been accumulated by DISISS and contains about 5,000 titles.

First attempts at clustering will be made using i) and iii) combined, as the set of source journals. It is expected that the results using this set will be judged more satisfactory than those using the random sample. This is partly because the random sample contains a number of journals with few or no citations, and partly because there is no guarantee that any subject area is adequately represented. It seems likely that the results using i) and iii) as sources will be more reliable that those using ii), in the sense that a similarly chosen sample will probably give similar results.

The data elements required for clustering are source identification and cited title for every citation in a journal. Each title must be identified by a unique code. Identification of source journals is simple, since codes can be given to the source titles listed in Appendix A. For cited titles the problem is less trivial since a title may be cited in several forms e.g. J. Exp. Psych., J. of Experimental Psychology, J. Exp. Psy. Since each title in CLOSSS is allocated a unique 5-digit number for book-keeping purposes, it was decided to use the CLOSSS number as the identification code for both source and cited journals. Cited titles not in CLOSSS (i.e. non-social science journals) are given a code consisting of 1 letter followed by 4 digits (a dummy CLOSSS number). Allocation of CLOSSS numbers cannot be achieved fully automatically as visual recognition of title variants is necessary to some extent. However the process only involves a small amount of manual checking and coding since only one occurrence of each title variant requires visual checking.

After allocation of CLOSSS numbers to cited titles, a simple counting procedure is required to produce the basic data matrix. This matrix consists of a row for each cited title, in which the elements are the number of citations from each of the source journals.

4.4 Treatment of data before clustering

The raw data consists of observations of the number of citations from each of the source journals to each journal title which is cited. Consider this as a matrix in which each row is a data point, representing a cited journal by the number of citations to it from each of the source journals. The number of columns of

this matrix is the number of source journals (variables) and the number of rows is the number of journal titles cited (data points).

4.4.1 Reduction in size of the data matrix

(a) Reduction in the number of rows

The time taken by the clustering algorithm increases rapidly with the number of data points. Since the journals which are seldom cited are not important in the context of clustering by citation pattern, the first step is to drop all journals which are cited only once. The pilot study data and a small sample of the ISI data indicate that this procedure could be expected to halve the number of data points. Extension of this process, dropping journals cited only twice or three times, is clearly possible. A further reduction in the number of data points can be made by omitting from the clustering data all titles cited by only one source journal. For our purposes such a journal must logically belong to the same clusters as the journal which cites it and can be added manually. There is obviously a possibility of a source journal not occurring or being dropped as a cited title. The second possibility can be avoided by artificially retaining a source journal as a cited journal if it would otherwise have been dropped. It is also possible to allocate titles manually to their nearest cluster, without affecting the degree of optimality of the clustering seriously.

(b) Reduction in the number of columns

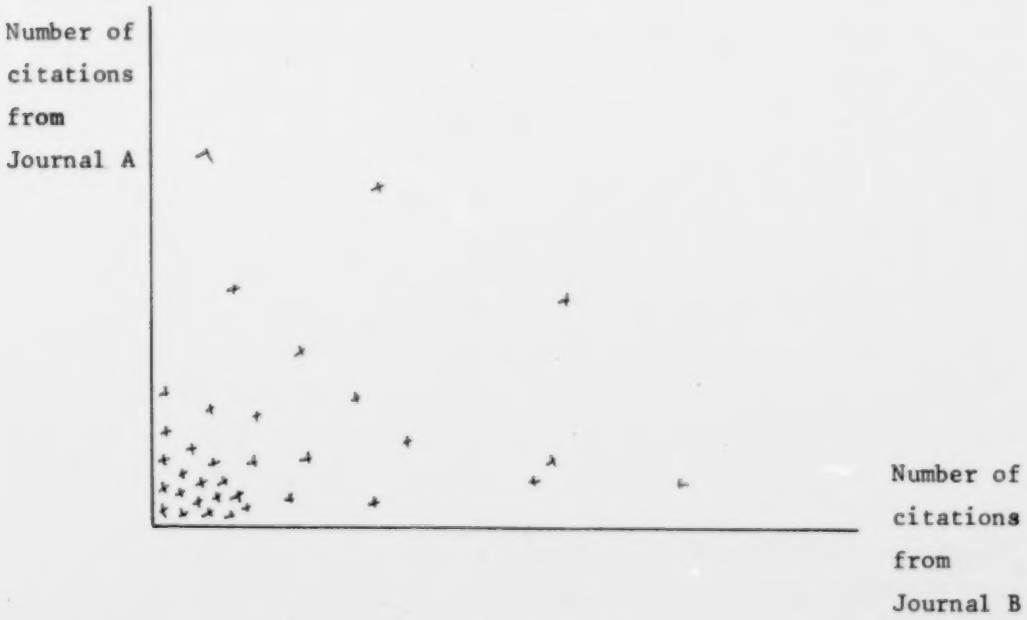
One approach would be to ignore columns with small entry totals (i.e. source journals which give few citations). A slight risk is run of failing to find small clusters concentrated on the ignored source journals.

Another approach is to analyse the data by principal component analysis prior to clustering. This procedure successively identifies linear combinations of the original columns which account for the maximum variance not explained by the previous linear combinations. Since maximum variance will tend to give maximum cluster differentiation it might be possible to select a somewhat smaller number of these linear combinations (principal components) as variables while still retaining most of the variance in the data as represented by the original variables. The variables will no

longer be identified with individual source journals but with particular linear combinations of the citations from these journals.

4.4.2 Normalization, self-citation adjustment and scaling

For a number of reasons the raw data is unsuitable for clustering directly. The dominant effect is that the data points tend to be allocated in the Euclidean space as in the diagram below, which is an attempt to represent a space of many dimensions on a two-dimensional sheet of paper.



The outlying points are the highly cited titles and the group in the centre are the remainder. The clustering algorithm would produce one very large cluster and a number of small ones, often consisting of a single journal[1]. Although this is reasonable given the data, it is not the sort of cluster pattern which is required. For our purposes a journal which is cited 10 times by A and 5 times by B would reasonably fall into the same cluster as a journal cited 100 times by A and 50 times by B. It would therefore be logical to take the proportion of the citations to a cited journal from each of the source journals as

[1]When 105 of the pilot study cited titles were analysed into 15 clusters using the raw data, 75 journals were grouped together and 8 of the other clusters consisted of a single journal.

the data. This is effected by dividing each matrix element by the total of the original elements in its row, and can be thought of as a form of row normalization.

Adjustment for self-citations can be made by reducing the matrix elements corresponding to self-citations, either by a fixed percentage, say 25, or by a percentage determined for the journal in question. Because the number of source journals is small compared with the number of journals to be clustered the effect of such adjustments will be very slight on non-source journals. Another form of adjustment which should be considered is scaling of the variables. This can have a large effect on the clustering as shown by the diagram below.

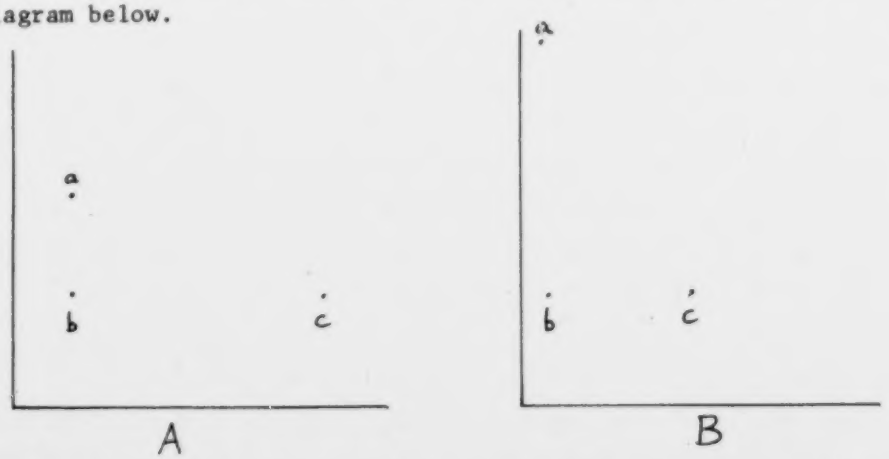


Diagram B is obtained from Diagram A by stretching the vertical axis and compressing the horizontal axis or taking $x' = \tilde{a}x$ ($\tilde{a} < 1$) and $y' = \tilde{b}y$ ($\tilde{b} > 1$). This scaling of the variables however alters the cluster pattern. In Diagram A, a and b would cluster together at an earlier level than c would cluster with either; but in Diagram B, b and c would cluster before a would cluster with either.

Two simple scaling procedures which could be applied are

(i) scaling by variable totals

(ii) scaling by standard deviation.

The first scales a column down in proportion to the number of citations from the relevant source journal, thereby increasing the importance of source journals which give few citations. In effect it says that if the same number of citations had been collected from each source

this is how they would have been distributed, extrapolating directly from the numbers that were collected.

The second procedure gives each column of the scaled matrix the same sample variance. The effect of this on the original correlated variables is somewhat indeterminate in general. However it might be a reasonable approach on the uncorrelated variables obtained by principal component analysis with the total variances on individual variables (off-diagonal entries of the covariance matrix being zero).

4.4.3 Choice of data adjustments

Apart from the row normalization, which is essential to our purpose, the only way to select which adjustments to make is by experiment. The pilot study data has been used for this purpose, and although the different size of the complete citation sample may demand a slightly different choice of adjustments, experience with the pilot study data will reduce the experimentation needed on the complete sample, or subsets of it.

4.5 Reliability and stability of clusters

If the results of clustering journal titles are to be used in the design of secondary services it is essential to assess the reliability of the clusters. There appear to be three main factors which may affect the clusters:

(i) date of source journals
(ii) selection of source journals
(iii) clustering method.

The stability of clusters over time is best studied by collecting citations from the same source journals from a number of different years and comparing the clusterings which are obtained using each year's data separately. It is obviously not feasible to collect citations from a large number of journals in a large number of years, but it is hoped to collect data from some criminology journals for 1950, 1960 and 1970.

Narin and Carpenter (1972) for scientific journals and Xhignesse and Osgood (1967) for psychology journals have both found a considerable amount of stability in citation structure over time, although it is clearly affected by the emergence of new journals and the disappearance of old ones.

The selection of source journals for the citation collection can clearly affect the frequency of citation of cited journals and thus the clustering results. Such an effect is obviously greater if only a few source journals are used. The number in the main citation collection is large by the usual standards of citation studies and should be sufficient to give reliable clusters. It is impossible to produce an objective criterion to decide which of two clusterings is better (except in a statistical sense which ceases to be applicable if different data is used). Clearly different sets of source journals may produce very different clusterings but only subjective judgement can be used to select one as superior, just as subjective judgement has to be used in the selection of source journals. It would certainly be reassuring if clustering using the randomly chosen source journals gave results similar to those obtained with the source journals in the ranked list from the pilot study. However, if this is not the case, it is no proof that the clusters using either list are in some way invalid. Intuitively, journals which are reasonably distributed over the social science disciplines and which are known to be fairly important should constitute a more satisfactory set of sources than any particular random sample of titles. A more useful test for the reliability of the chosen set of sources might be to divide the ranked list of journals into two sections trying as far as possible to maintain the distribution over discipline, language, etc. If these two sets of source journals produced results similar to each other and to those obtained using the combined set, then a great deal of confidence could be placed in the reliability of the results in so far as selection of source journals was concerned.

The effect of particular clustering methods on the clusters obtained from a given set of data has been discussed in passing in section 4.2.3. As mentioned there, Narin and Carpenter (1972) made the assumption that clusters which were found by several methods were

reliable clusters. The results using the pilot study data with the SCICON method and single- and complete-linkage (see sections 5.3 and 5.4) suggest that the main clusters identified by the SCICON method were also detected by the other approaches. It is also necessary to consider a sequence of results obtained by the SCICON method for different numbers of clusters, to get an idea of the stability of clusters within the process. A cluster at one level which distributes itself between several different clusters when the number of clusters is reduced is clearly not stable even within the data collected and the clustering process used. It is worth stating again that the SCICON criterion tends not to produce clusters of very different sizes at any one level. This means that the effect of outlying points may be large and omissions of obvious outliers from the clustering runs should be considered.

4.6 Evaluation and representation of clusters

Evaluation of clusters is possible only within the context of the use to which the results are to be put. Hence clusters of articles used in the design of an information retrieval system can be evaluated only by testing them with user requests, either in a real system over a period of time by monitoring user reaction, or by simulating the results using selected requests for which recall and relevance ratios (or some similar concepts) can be measured.

It is not immediately obvious how clusters of journals intended as a basis for secondary service coverage can be evaluated. Comparison with conventional subject classification would be interesting, but if there are substantial differences a decision (which implies an evaluation) is necessary as to which breakdown is more suitable for selecting journals to cover in a particular secondary service.[1] An estimate of the proportion of citations from journals within the cluster to journals outside that cluster would be useful.

Associated with the problem of evaluating clusters is that of representing them in such a way that the maximum amount of useful information is extracted from the results, bearing in mind the aims of the clustering. Long lists of journals which form the basic output from

[1] Since the evaluation would be subjective, and conventional classifications

the computer program are cumbersome and impossible to assimilate without some condensing and subsidiary analysis. Diagrammatic representation is difficult as the Euclidean space in which the clusters are embedded is of more than two dimensions. However it is usually easier to get information from diagrams than from tables of figures, and some effort in this direction might be well repaid. Some attempts at compressing the results of clustering using the pilot study data are made in section 5.3.

5 ANALYSIS OF THE PILOT STUDY DATA

During the summer of 1971, citation data was collected from 17 source journals in the main subject areas of the social sciences (see Appendix B). This data has been discussed in Working Paper No. 5 from a descriptive point of view. This data has been used to test the program for the SCICON algorithm for programming errors and to investigate the performance of the algorithm on this type of data. The effects of various treatments of the data on the clusters produced have also been studied. In addition some of the data has been used with three other clustering algorithms:

i) single-linkage hierarchical clustering;
ii) complete-linkage hierarchical clustering;
iii) an algorithm suggested by C.J. Van Rijsbergen (1970).

These algorithms were selected for simplicity of application, and, because they represent a variety of approaches, comparison of the results obtained with those from the SCICON method is of interest.

5.1 Modification to clustering program for SCICON method

The program being used was originally programmed for the ICL 1900 series computers by David Hitchin of the University of Sussex in a combination of PLAN and FORTRAN. Some later modifications have been made to cater for larger amounts of input data. These are:

i) a paging facility has been added to allow part of the data to be held on disc rather than in core storage while the program is running;

ii) a time-consuming and widely criticised measure of cluster precision has been omitted.

5.2 Description of the pilot study data

The data used for clustering was for all journals cited more than once in the pilot study, using as sources the journals listed in Appendix A. The data is listed in Appendix B. Those titles marked with an asterisk were omitted from the clustering runs because they were cited by only one source journal. Such titles must logically be in the same cluster as the relevant source journal and manual addition can be made after the runs. The number of titles in the clustering runs is 115, with a further 98 being cited by only one source.

Clustering runs were made using several treatments of the data:

i) unmodified, as in Appendix B;
ii) with cells divided by row and column totals;
iii) with cells divided by row totals only;
iv) with cells divided by row totals and a constant added to each non-zero cell;
v) as iii) with self-citation cells reduced by 25% prior to division by row totals;
vi) as iii) with self-citation cells which were the highest cell in their row reduced to the next highest cell in the row, this adjustment being made prior to division by row totals.

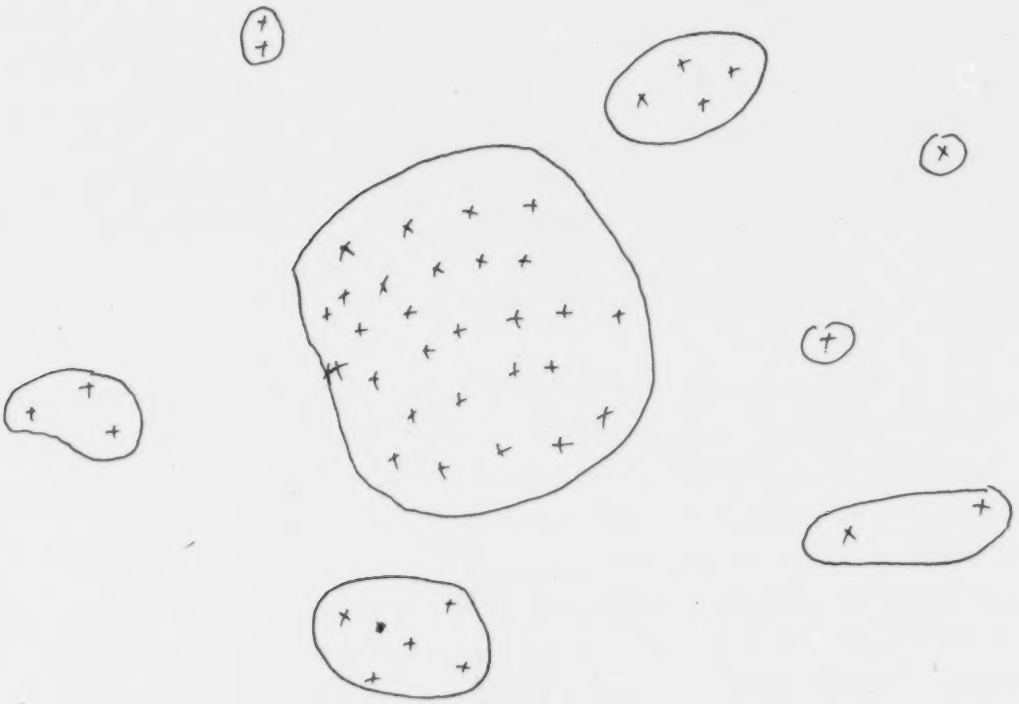
5.3 Results using the SCICON algorithm

Brief results are given here for most of the data treatments listed in section 5.2, together with more detailed results for the most promising treatments.

i) Unmodified data

This run demonstrated the necessity for treating the data in some of the ways suggested in section 4.4. When the journals were grouped into 10 clusters, 80 of the 115 journals appeared in one cluster centered near the origin (the typical journal of this cluster would be cited between .03 and .91 times by each of the sources).

The remainder were the most highly cited journals in the sample, including 11 of the 17 source journals. The data and the clusters produced are of the form shown in the diagram below.



The outlying points represent the highly cited titles which have high values in one or more variables.

ii) With cells divided by row and column totals

It seems reasonable to adjust for differences in the number of citations collected from each source to approximate the situation where the same number of citations are collected from each source. However, from the run it appeared that this procedure gave too much weight to citations from journals from which fewer citations were collected. Since on the whole these journals were the less

important ones, it seemed unreasonable that they should function as discriminators between clusters rather than the more important titles which provided more citations.

In theory it would seem reasonable to adopt one of two strategies when collecting citations for clustering:

a) collect strictly equal numbers of citations from each source journal;
b) collect all citations (or every nth) from all articles (or every mth) for a year, and use the implicit weighting as a measure of the importance of a source for discriminating betweeen clusters.

The strategies will clearly produce different clusterings but at least underlying properties of the data are known and results can be considered in the light of these features. A possible argument against the second is that journals have widely different numbers of citations per article. In the pilot study the second strategy was attempted, using every third article for each of 1950, 1960 and 1970. Unfortunately some of the sources had not been in existence long enough, and, for some, data collection could not be completed in the time allowed. Despite these shortcomings it is probably better not to adjust for the different numbers of citations from the source journals, but to allow the implicit weighting to have an effect on the clusters.

iii) With cells divided by row totals only.

More extensive runs have been performed on this form of the adjusted data than for the others. The 115 journal titles have been grouped into numbers of clusters in the range 25 to 3. A problem which requires attention is the representation of these results (see section 4.5). An attempt is made here to give useful and descriptive representations but inevitably each representation ignores some of the information available from the results.

One approach is to look at the individual journals and see which cluster they are in at different levels. In Appendix E, the journals in each of the clusters at the 3-cluster levels are traced through some of the cluster levels (the levels were chosen arbitrarily). Of these 3 clusters, one is clearly a psychology cluster of 34 journals and another an economics cluster of 21 journals. The third cluster is less easy to classify and contains 60 journals. If we consider the journal traces (see Appendix E) of the first two clusters, it is clear that most journals have one of a small number of traces. Giving these traces codes we can group the journals according to their trace codes as in Tables 5.1 and 5.2. The traces can be used to display a simple network for each of these clusters.

Fig 5.1

Psychology Cluster

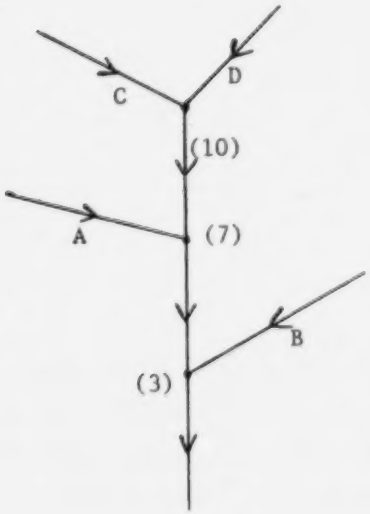
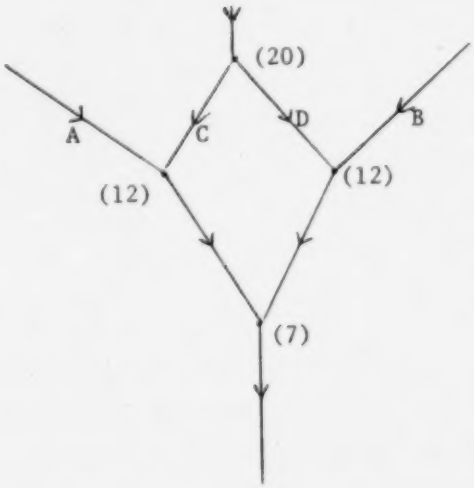


Fig 5.2

Economics Cluster



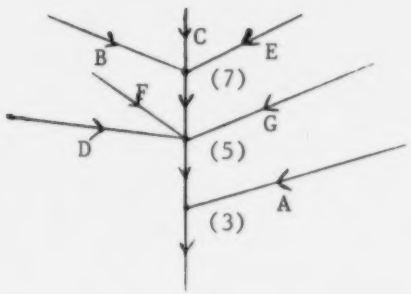
The levels attached to these diagrams have been taken from the traces. More accurate, though not necessarily more helpful, levels could be obtained from the original clustering print-out. Comparison of levels gives a rough guide to the relative cohesion of clusters.

For example, the economics cluster is a separate entity from level 7 downwards and is therefore more cohesive than the psychology cluster, which is in two sub-clusters at level 7 which do not combine till level 3.

Unfortunately the third, rather amorphous, cluster does not lend itself so easily to this type of analysis. By restricting the trace to level 10 and below a good proportion of the journals can be grouped by trace code (see Table 5.3) and the corresponding network is shown below. It must be remembered that the structure of this network is less well-defined than the previous two because of the restriction on the trace codes .

Fig 5.3

Amorphous Cluster



Considering these subgroupings subjectively it is interesting to try and identify the common features of journals in a sub-cluster. In the psychology cluster, A might be called 'general psychology'; B includes all the German psychology journals; C is perhaps 'cure and prevention of psychological disturbance'; and D is too small to merit a name but is close to C. In the economics cluster: A seems general; B is clearly 'political and social economics'; and C and D are biassed towards 'economic statistics and statistical methods'. Even in the amorphous cluster subjective classification reveals some structure: A is probably 'social anthropology'; B 'sociology'; C is clearly related to the psychology group C[1] and is to do with

---

[1] In fact, reference to the original clustering shows that psychology groups C and D and group C from the amorphous cluster, together with J.Experimental Pyschology, constitute cluster 4 at level 10.

curing and preventing psychological and social disturbances; D is rather political; F is definitely French; and E and G might be thought of as less academic journals.

Another way of looking at the clustering results is to start from the point of view of the clusters rather than that of the journals in them. An attempt can also be made to follow the clusters obtained at one level, as the number of clusters is reduced. As an example the clustering into 8 clusters can be briefly described as follows.

Cluster no.	Size	Mean distance from cluster centre	Description
1	25	5.613	Sociology & Social Psychiatry
2	4	3.388	French Economics
3	19	4.869	Social Anthropology
4	8	4.245	Politics
5	25	3.320	General Psychology
6	10	3.752	Economic statistics
7	12	3.744	Psychology (German plus oddments)
8	12	2.916	Economics

The mean distance from the cluster centre of the items in the cluster is a measure of the compactness of the cluster. Thus cluster 5 is much more compact than cluster 1 even though the number of items in each is the same. As the number of clusters is reduced the clusters merge and lose and gain items as follows.

7 clusters

The economics clusters 6 and 8 merge to form 6, losing one item to 5 and two to 3. One item moves from 3 to 1.

6 clusters

The French economics cluster 2 is absorbed into cluster 1, which also swaps several items with cluster 3 and gains one item each from 7 and 6, leaving the situation below.

Cluster no.	Size	Mean distance from cluster centre	Description
1	33	6.225	Sociology, Social Psychiatry and French Economics
3	19	4.065	Social Anthropology
4	8	4.245	Politics
5	26	3.404	General Psychology
6	18	4.153	Economics
7	11	3.507	Psychology (German + oddments)

5 clusters

Cluster 4 is absorbed into cluster 1 and one item moves back from cluster 3 into the enlarged cluster 1.

4 clusters

Clusters 1 and 3 combine and lose three items to the economics cluster 6.

3 clusters

The two psychology clusters 5 and 7 merge, losing three items to the large cluster 1.

From this,it can be seen that the general psychology cluster 5 is fairly isolated and remains practically unchanged until finally merged with 7,which is also relatively isolated but much less dense. The economics clusters 6 and 8 merge to form a slightly less isolated cluster 6,with a few points lying between it and cluster 1. There is clearly some overlap between clusters 1 and 3,until they merge to form a large cluster which is not very compact.

Although the mathematical and computational problems of allowing overlapping clusters are considerable, it is worth considering the output from the SCICON program which might be relevant to lifting this restriction. For each clustering ,the distance of each point from its own cluster, its distance from the next nearest cluster and the ratio of these distances are output. Looking at the three clusters listed in Appendix E, of the 60 points in the amorphous cluster all but 5 are closer to the economics cluster than to the psychology cluster. All the points in the economics cluster and all but one in the psychology cluster have the amorphous cluster as their second nearest cluster. If we consider all journals with the ratio

$$\frac{\text{distance to next nearest cluster}}{\text{distance to own cluster centre}} > 1.15$$

as points overlap, we get slightly over 10% of points in more than one cluster (Table 5.4).

Table 5.4

Points of overlap

Clusters	Economics	Amorphous	Psychology
Mean distance from cluster centre	4.33	6.31	4.12
056	Cambridge Journal (7.4)		
059	Can. J. Econ. & Pol. Sci. (.48)		
137	J. Opt. Soc. Am. (5.7)		
163	Parliamentary Affairs (10.1)		
193	Revue d'Econ. Politique (9.1)		
209	Social Service Quarterly (10.2)		
223	Yale Law Journal (6.4)		
		018	Am. Psychologist (7.6)
		116	J. Educ. Psychology (6.9)
		161	Oxford Economic Papers (8.3)*
		173	Psychol. Arbeiten (8.3)
		180	Psychologische Forschung (7.8)
		202	Science (5.5)

Note

The figues in brackets are the distances of the points from their cluster centre.

*A miscoding of the data for this journal results in its acting like a psychology journal (see Appendix B).

Clearly some of these points can genuinely be considered as belonging to two clusters (059 and 202) while others do not fit at all well into either (163 and 209). If it were required to determine which journals should be covered by two secondary services it would clearly be necessary to eliminate the latter type, either by subjective judgement or by some objective criterion. For instance, it could be required that the distance of a genuine overlap point from its secondary cluster centre be less than one and a half times the mean distance of points in that cluster from the centre.

With this criterion only those journals underlined would qualify as overlap points. From this example it would seem that this criterion favours the selection of points from compact clusters as overlap points rather than those from less compact ones.

This description of the results of clustering the data with each cell divided by its row total is rather long but is intended to highlight the problems of representing the clusters in the most helpful form for subsequent use,as discussed in section 4.5.

(iv) With cells divided by row totals and a constant added to each cell

Behind this adjustment is the idea that for a journal to be cited at all by a source journal could be considered as more important, in differentiating between it and another journal not cited by the source, than a small difference in the percentage of citations to each of the journals coming from a particular source. It was hoped that this process might reveal more structure in the amorphous cluster. This was not in fact borne out,and the clusterings obtained from 10 to 3 clusters were broadly similar to the results without the addition of the constant (regarding the values of the variables for a journal as the percentage of citations to that journal which came from each source, 10% was added to each value). Since there is no best clustering it is difficult to decide which of two is to be preferred when neither has any obvious relative merit or disadvantage. Although the detailed composition of the clusters obtained differs slightly, the general description of the clusters revealed in this case and in case (iii) are the same. At the 3-cluster level the only differences were that the economics cluster gained three French economics journals from the amorphous cluster, and the psychology cluster gained American Psychologist and Science, compared with case (iii).

(v) With self-citations reduced by 25% before dividing by row totals

The effect of this adjustment was negligible at all levels between 10 and 4 clusters, and the clusters at the 3-cluster level were identical with case (iii).

(vi) With self-citation cells which were the highest in their row reduced to the level of the next highest cell before dividing by row totals

Because the extent of self-citation varies considerably between journals the approach used in (v) has been criticized. A different approach which has maximum effect on journals with a high rate of self-citation was therefore tried. The effect was small but mainly on those source journals which had a high rate of self-citation. At the 3-cluster level American Psychologist moved into the psychology cluster, and Economica and Psychologische Forschung moved nearer the centres of the economics and psychology clusters respectively.

5.4 Results using other approaches

To compare results given by the SCICON algorithm with those of some other approaches, the data divided by row totals (case (iii) in section 5.3) for the 34 journals in the psychology cluster together with American Psychologist and Science was used with three other algorithms. Those algorithms were chosen as being simple to apply without excessive programming effort. The algorithm suggested by C.J. Van Rijsbergen (1970) appeared as a FORTRAN program in Computer Journal and required only the incorporation of calculation of a similarity matrix. Euclidean distance was used for this purpose. Intermediate output from this program included a listing of the similarity elements sorted in increasing magnitude. This listing was used for hand-calculation of the single- and complete-linkage clusterings.

Van Rijsbergen's algorithm identifies all clusters having the property that the maximum distance between any pair of points in the cluster is less than the minimum distance of any point in the cluster to any point not in the cluster. The clusters which satisfy this very stringent condition are listed in Appendix F. Of these, the first three are inevitable, since the variable values for the journals in each cluster are identical. The nature of the condition for a cluster ensures that any data is bound to contain at least one pair of points satisfying the condition, and not much importance can be attached to two point clusters. These results merely show that this simple but stringent condition is of no use to DISISS since the data does not form very distinct isolated clusters.

The dendrograms obtained from the single- and complete-linkage algorithms are displayed in Figures 5.5 and 5.6. The circled groups identify the tight clusters found by Van Rijsbergen's algorithm. Also added to the figures are the trace codes obtained for the journals in the psychology cluster in section 5.3 (see Table 5.1). The following facts should be noted when considering the dendrograms:-

(i) The exact ordering of the journals is irrelevant. The hierarchy could have been displayed using a large number of different permutations of the journals. For instance, the sequence of the journals in Figure 5.5 between J. Abnormal and Social Psychology and Psychologische Forschung could have been placed before Acta Psychologica.

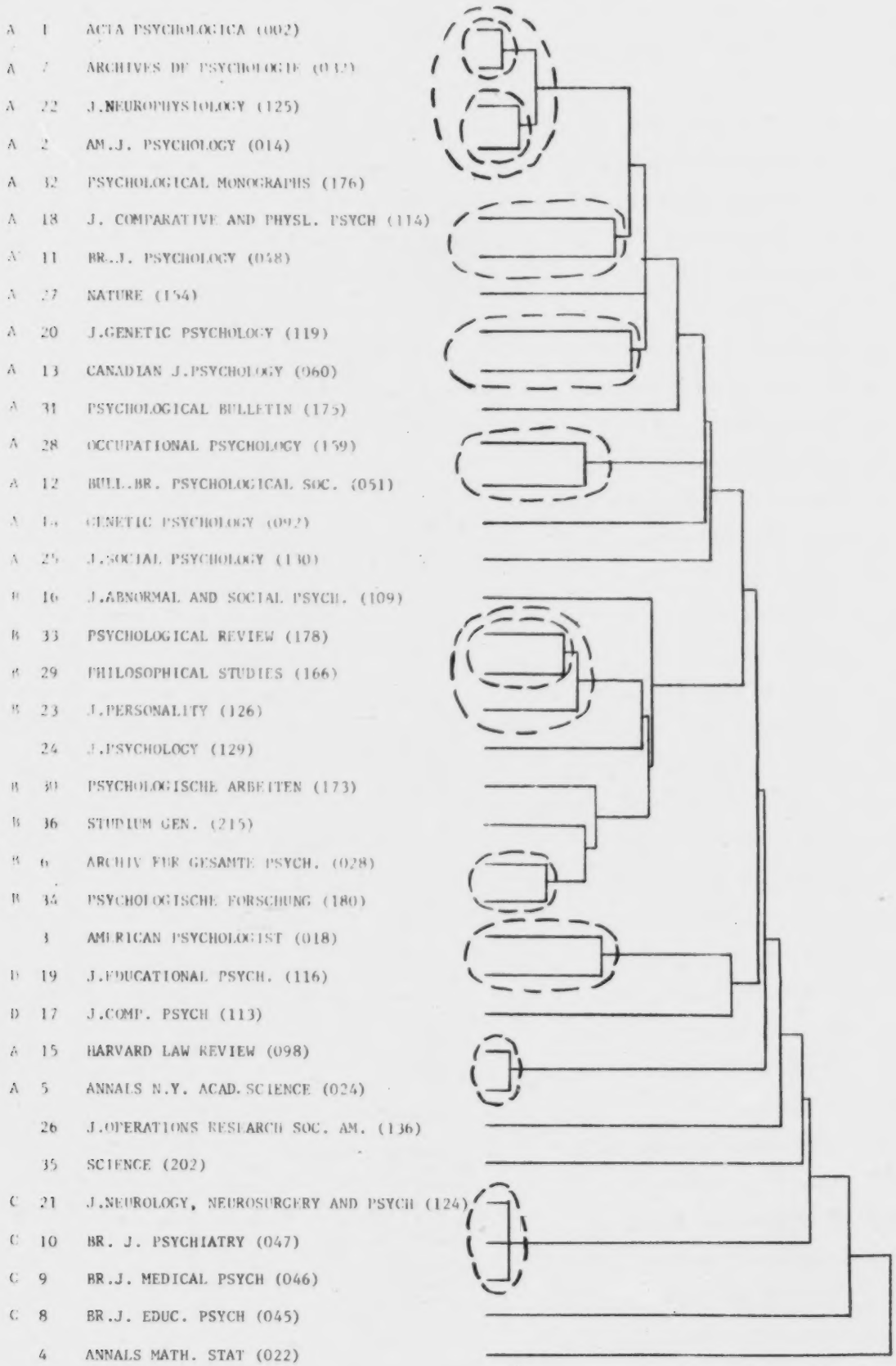
(ii) In Figure 5.5 (single-linkage) the level[1] at which two sub-clusters combine to form one cluster is the minimum distance between any point in one sub-cluster and any point in the other.

(iii) In Figure 5.6 (complete-linkage) the level at which two sub-clusters combine is the maximum distance between any pair of points in the combined cluster.

Both methods clearly identify the journals with trace code B (and J. Psychology, which lay between A and B) as a cluster at a relatively low level, and at a similar level a large proportion of the A journals are clustered. The exact behaviour of the more peripheral (in terms of clustering rather than of subject matter) journals differs, the difference being a function of the particular algorithms (see section 4.2.3). The complete-linkage dendrogram also identifies the C and D groups reasonably well and produces the type of clustering which could be of use to DISISS. The 'chaining' tendency of the single-linkage algorithm tends to obscure the structure of Figure 5.1, which could be very useful in the design of secondary services.

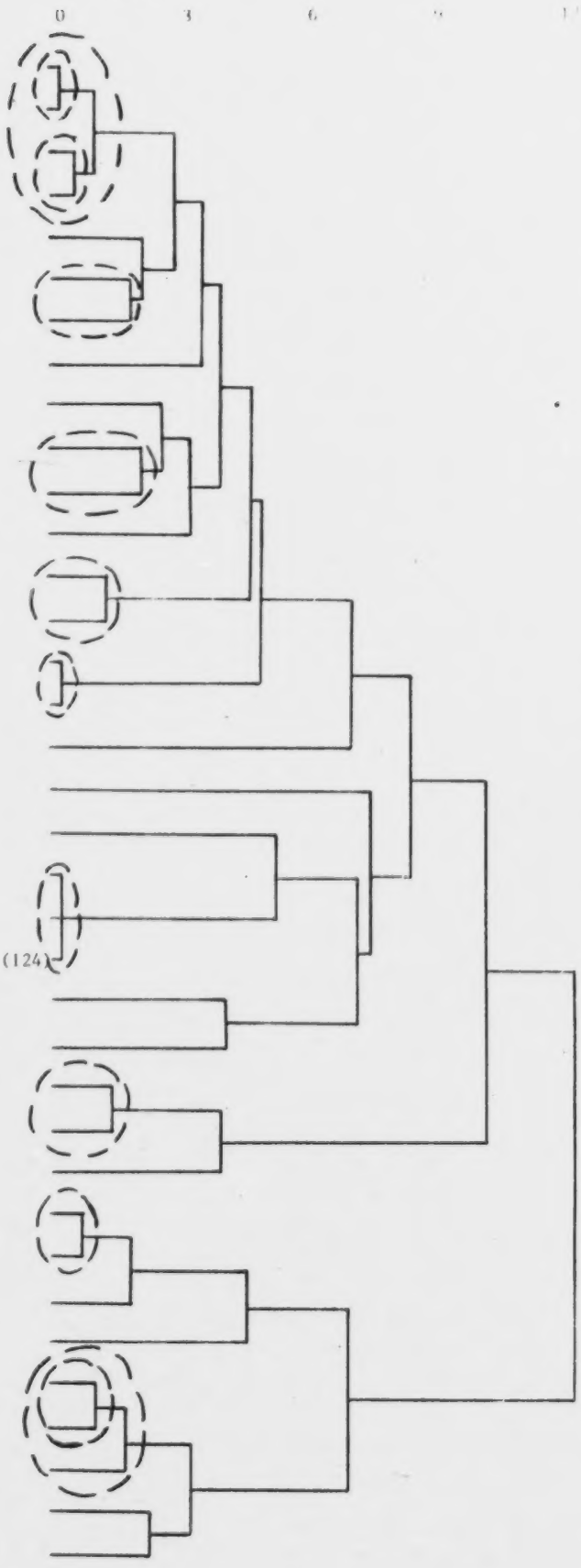
---

[1] The use of the term 'level' in this context should not be confused with the level of clustering referred to with the SCICON method, which is the number of clusters.


A 1 ACTA PSYCHOLOGICA (002)
A 7 ARCHIVES DE PSYCHOLOGIE (013)
A 22 J.NEUROPHYSIOLOGY (125)
A 2 AM.J. PSYCHOLOGY (014)
A 32 PSYCHOLOGICAL MONOGRAPHS (176)
A 18 J. COMPARATIVE AND PHYSL. PSYCH (114)
A 11 BR.J. PSYCHOLOGY (048)
A 27 NATURE (154)
A 20 J.GENETIC PSYCHOLOGY (119)
A 13 CANADIAN J.PSYCHOLOGY (060)
A 31 PSYCHOLOGICAL BULLETIN (175)
A 28 OCCUPATIONAL PSYCHOLOGY (139)
A 12 BULL.BR. PSYCHOLOGICAL SOC. (051)
A 14 GENETIC PSYCHOLOGY (092)
A 25 J.SOCIAL PSYCHOLOGY (130)
B 16 J.ABNORMAL AND SOCIAL PSYCH. (109)
B 33 PSYCHOLOGICAL REVIEW (178)
B 29 PHILOSOPHICAL STUDIES (166)
B 23 J.PERSONALITY (126)
24 J.PSYCHOLOGY (129)
B 30 PSYCHOLOGISCHE ARBEITEN (173)
B 36 STUDIUM GEN. (215)
B 6 ARCHIV FUR GESAMTE PSYCH. (028)
B 34 PSYCHOLOGISCHE FORSCHUNG (180)
3 AMERICAN PSYCHOLOGIST (018)
D 19 J.EDUCATIONAL PSYCH. (116)
D 17 J.COMP. PSYCH (113)
A 15 HARVARD LAW REVIEW (098)
A 5 ANNALS N.Y. ACAD. SCIENCE (024)
26 J.OPERATIONS RESEARCH SOC. AM. (136)
35 SCIENCE (202)
C 21 J.NEUROLOGY, NEUROSURGERY AND PSYCH (124)
C 10 BR. J. PSYCHIATRY (047)
C 9 BR.J. MEDICAL PSYCH (046)
C 8 BR.J. EDUC. PSYCH (045)
4 ANNALS MATH. STAT (022)

SINGLE LINEAGE HIERARCHICAL CLUSTERING OF PSYCHOLOGY CLUSTER

Fig 5.5


0
3
6
9
12
A 1 ACTA PSYCHOLOGICA (002)
A 7 ARCHIVES DE PSYCHOLOGIE (032)
A 22 J. NEUROPHYSIOLOGY (125)
A 2 AM.J. PSYCHOLOGY (014)
A 32 PSYCHOLOGICAL MONOGRAPHS (176)
A 18 J.COMPARATIVE AND PHYSL.PSYCH (114)
A 11 BR.J. PSYCHOLOGY (048)
A 14 GENETIC PSYCHOLOGY (092)
A 27 NATURE (154)
A 20 J.GENETIC PSYCHOLOGY (119)
A 13 CANADIAN J. PSYCHOLOGY (060)
A 31 PSYCHOLOGICAL BULLETIN (175)
A 28 OCCUPATIONAL PSYCHOLOGY (159)
A 12 BULL.BR.PSYCHOLOGICAL SOC.(051)
A 5 ANNALS N.Y.ACAD SCIENCE (024)
A 15 HARVARD LAW REVIEW (098)
4 ANNALS MATH. STAT. (022)
35 SCIENCE (202)
C 8 BR.J.EDUC.PSYCH. (045)
C 9 BR.J.MEDICAL PSYCH. (046)
C 10 BR.J.PSYCHIATRY (047)
C 21 J.NEUROLOGY, NEUROSURGERY AND PSYCH. (124)
A 25 J.SOCIAL PSYCHOLOGY (130)
26 J.OPERATIONS RESEARCH SOC. AM. (136)
3 AMERICAN PSYCHOLOGIST (018)
D 19 J.EDUCATIONAL PSYCH. (116)
D 17 J.COMPARATIVE PSYCH. (113)
B 6 ARCHIV FOR GESAMTE PSYCH. (028)
B 34 PSYCHOLOGISCHE FORSCHUNG (180)
B 36 STUDIUM GEN. (215)
B 30 PSYCHOLOGISCHE ARBEITEN (173)
B 29 PHILOSOPHICAL STUDIES (166)
B 33 PSYCHOLOGICAL REVIEW (178)
B 23 J.PERSONALITY (126)
B 16 J.ABNORMAL AND SOCIAL PSYCH. (109)
24 J.PSYCHOLOGY (129)

COMPLETE LINKAGE HIERARCHICAL CLUSTERING OF PSYCHOLOGY CLUSTER

Fig 5.6

5.5 CONCLUSIONS

(i) Since the single and complete-linkage algorithms represent extremes of their family of hierarchical methods, it is encouraging that they and the SCICON method identify the B clusters as distinct from the larger A group, and,less importantly, detect the C and D groups. This implies that the structure revealed by the SCICON method is in the data and not imposed by the algorithm.

(ii) Van Rijsbergen's algorithm and single-linkage can be discarded as not providing results of use to DISISS.

(iii) The complete-linkage algorithm provides useful results but is computationally infeasible for large amounts of data (see section 4.2.3).

(iv) The most suitable treatment for the raw data appears to be to divide data cells by their row totals after reducing self-citation cells which are the highest in their row to the level of the next highest cell. Apart from self-citations this means that the data used is the percentage of citations to a journal which are found in each of the source journals. The possibility of adding a constant to each non-zero cell should not however be discarded.

(v) The source journals which appeared as cited titles but were cited only by themselves were included in the clustering runs but tended to be outliers distorting the clusters in the rest of the data. As mentioned in Section 4.2.3, the SCICON method is unlikely to produce clusters of very different sizes and thus outliers will not appear as single point clusters. For this reason it is probably better to exclude all titles cited by only one source. A related point is that the clustering obtained from the ranked list of journals should be more informative with less cited titles excluded for this reason.

Table 5.1

Psychology Cluster

Trace Code		
A	002	Acta Psychologica
	014	Am.J. Psychology
	024	Annals N.Y. Acad.Sci.
	032	Archives de Psychologie
	048	Br. J.Psychology
	051	Bull. Br.Psychological Soc.
	060	Canadian J.Psychology
	092	Genetic Psychology Monographs
	098	Harvard Law Review
	114	J.Comp. and Physl. Psych.
	119	J.Genetic Psychology
	125	J.Neurophysiology
	130	J.Social Psychology
	154	Nature
	159	Occupational Psychology
	175	Psychological Bulletin
	176	Psychological Monographs
B	028	Arch.für Gesamte Psychologie
	109	J.Abnormal and Social Psychiatry
	126	J.Personality
	166	Philosophical Studies
	173	Psychologische Arbeiten
	178	Psychological Review
	180	Psychologische Forschung
	215	Studium Gen.
C	045	Br.J.Educational Psychology
	046	Br.J.Medical Psychology
	047	Br.J.Psychiatry
	124	J.Neurology, Neurosurgery and Psychiatry
D	113	J.Comparative Psychology
	116	J.Educational Psychology

Table 5.1 (continued)

+	022	Annals of Mathematical Statistics
	129	J.Psychology (between A and B)
	136	J.Operations Research Soc.Am. (near A)

Table 5.2

Economics Cluster

A	011	Am.Economic Review
	080	Ekonomisk Tidsskrift
	153	National Institute Econ. Review
	188	Quarterly J.Economics
	191	Review of Economic Studies
B	023	Annals Am.Acad.Pol. and Social Sci.
	031	Arch fur Sozialwissenschaft u.s.w.
	037	Australian Quarterly
	151	Monthly Labour Review
	216	Survey of Current Business
C	052	Bull. Oxford Inst.Statistics
	072	Econometrica
	073	Economic Journal
	128	J.Political Economy
	138	J.Royal Statistical Society
D	076	Economica
	190	Rev.Economics and Statistics
+	059	Canadian J.Econ. and Pol Sci. (almost A)
	075	Economic Weekly (Bombay) (almost B)
	137	J.Opt.Society of America
	147	Manchester School

Table 5.3

Amorphous Cluster

Trace Code

A	007	Africa
	008	Am.Anthropologist
	050	Br.J.Sociology
	066	Colliery Guardian
	100	Human Organization
	101	Human Relations
	127	J.Philosophy
	131	J.Social Issues
	162	Pacific Sociological Review
	170	Proc. Aristotelian Soc.
	212	Sociological Review
B	010	Am.Behavioral Scientist
	015	Am.J.Sociology
	019	Am.Sociological Review
	064	China Weekly Review
	069	Current Sociology
	093	Geographical Review
	185	Public Opinion Quarterly
	196	Rev.Int. de Sociologie
	211	Sociological Quarterly
	213	Sociometry
C	012	Am J.Orthopsychiatry
	013	Am J.Psychiatry
	018	Am.Psychologist
	043	Br.J.Criminology
	223	Yale Law Journal

Table 5.3 (continued)

D	004	Administrative Science Q.
	017	Am.Political Science Rev.
	036	Australian Outlook
	056	Cambridge Journal
	085	Esprit
	104	Industrial & Labour Relations Rev.
	157	New Statesman
	207	Social Forces
E	033	Comp.Studies in Society & History
	163	Parliamentary Affairs
	171	Psychiatry
	202	Science
	209	Social Service Quarterly
F	003	Act.Econ. et Financiere
	193	Rev. d'Economie Politique
	194	Rev. Economique
	195	Rev.Francaise de Sci.Pol.
G	025	Année Sociologique
	083	Encounter
	141	Kyklos
	189	Review
	222	World Politics
+	016	Am.Mus. of Nat. Hist. Anthrop. Papers
	021	Annals
	034	Aust.and N.Z. J.Sociology
	106	Int. J.Social Psychiatry (almost B)
	115	J.Crim.Law & Criminology (SI)
	118	J.Experimental Psych.
	135	J.Am.Stat.Association
	145	Listener
	161	Oxford Economic Papers
	208	Social Problems (almost B)
	217	Time

6 PROGRESS WITH DATA COLLECTION AND CONVERSION

6.1 ISI data

Magnetic tapes of Science Citation Index for one quarter of 1971 were made available to DISISS, and a copy of these tapes was taken from the SCI tapes held by the United Kingdom Chemical Information Service (UKCIS). The tapes required three types of conversion.

(a) The tapes required copying from 9 track tapes (as generally used on IBM or ICL System 4 machines) to 7 track tapes (for use on the ICL 1900 series).

(b) The six-bit character code BCD used by IBM and ICL System 4 had to be converted to the six-bit code used on ICL 1900 machines.

(c) The ISI record format had to be converted to the format used for citation data collected in the field by DISISS.

Stage (a) involved using the Bristol University ICL System 4-70 which is equipped with 7 and 9 track tape decks. A standard program was used which converted IBM EBCDIC characters to their BCD equivalents. Each 9 track tape was converted into two 7 track tapes with the following properties.

(a) 556 bits per inch.
(b) 0.75 inch interblock gap.
(c) Odd parity.

All further computer work has been carried out at the Open University on the ICL 1903A.

Stage (b) was accomplished by program using a character conversion table.

Stage (c) was combined with the process of extracting all records from the selected source journals and attaching CLOSSS[1] numbers to these journals.

At this stage, the ISI data is comparable with the data collected in the field, after punching, input and data vetting. The main problem to be overcome before clustering is to identify all citations to the same journal even if the forms of the title differ (e.g. with differing abbreviations), and to attach CLOSSS numbers to the cited titles. Two programs have been written to achieve this. When repeatedly applied to the data as it is converted and accumulated, they take advantage of as much automation as possible without sophisticated linguistic analysis of titles.

Program A takes as input a tape sorted by cited title in such a way that for each title version any occurrences with real CLOSSS numbers attached precede any with dummy CLOSSS numbers, which precede any with no CLOSSS numbers. It produces as output a tape on which all occurrences of a title version will be associated with a CLOSSS number, which will be real if a real number was attached to any occurrence of that version and dummy otherwise. A list of cited title versions with attached CLOSSS numbers and frequency of occurrence is also output.

This list is checked manually to produce card input data for program B, which replaces dummy CLOSSS numbers by real ones where relevant, and ties up different title versions for the same journal by giving them the same CLOSSS number (real or dummy). When cycling through programs A and B with an increasing volume of data, only new title versions require manual intervention on subsequent cycles. The output from program A can also be used to identify highly cited journals which are not currently in CLOSSS but should be considered for addition.

CLOSSS[1] (Check List of Social Science Serials) is being assembled as another part of the DISISS project. Since it is more convenient to identify journals by a unique fixed length code than by name, the numbers attached to the CLOSSS records are being added to citation records for all source and cited journals which occur in CLOSSS. Non-CLOSSS titles are given dummy CLOSSS numbers.

At 1st April 1973, the first 3 tapes of the original 7 tape file have been processed to this stage for the data required for clusterings. The last 4 tapes are being converted at Bristol.

The final step required to produce the clustering data is to count the citations to each title, as identified by its CLOSSS number, from each source journal, and to produce a tape containing this information.

6.2 Data collected in the field

Field collection of citation data from social science source journals took place in Summer 1971 for the pilot citation study and in Easter and Summer 1972 for the main serials citation file. The former data has been used for the development of clustering programs; the latter data, from the main file, will be used for the main clustering runs in conjunction with the data taken from SCI tapes. Some additional criminology data might also be used for clustering runs. The field collection of citation data was undertaken by researchers and students at the Polytechnic of North London School of Librarianship.

The pilot study file contained 4,918 citations. The main file will contain over 40,000 citations, at least 40% of which are to journals. The citations to monographs will not be clustered. The main file and the criminology file will both be used, in addition, for descriptive studies of citations.

6.3 Conversion of field collected data

Over 40,000 records from 120 source journals have been collected. A program has been written to vet the data, check sequences, check numeric data and to ensure that authors and titles are valid. The punching work was shared between the Computer Unit at Bath University and a bureau at Weston-super-Mare. Punching began in July 1972, and is virtually complete (April 1973). The creation of the main citation file was begun in January 1973, and by the middle of March, 20,000 records had been processed. Work is progressing on writing programs for the allocation of CLOSSS numbers to the cited journals titles, and it is hoped to complete and merge the DISISS file with the ISI/SCI file by August 1973.

7 FUTURE WORK

The next stage of the work is to obtain first clusterings using the main citation file. This involves the following.

(i) Completion of conversion of ISI data to DISISS format (see section 6.1).

(ii) Completion of the creation of the file of data collected in the field (see section 6.2).

(iii) Unification of cited journal titles and allocation of CLOSSS numbers to those titles for both sets of data.

(iv) Calculation of the basic data matrix required for clustering, i.e. counting the number of citations to each cited title from each source journal.

At this stage it will be necessary to consider the treatment of the data to be used for the first clustering runs. The types of source journals to include must be chosen, and also a cut-off level of citation below which cited titles will be omitted. The first run will probably include only citations from the ranked list of sources and use a high cut-off level to reduce the matrix size. As suggested in section 5.5, self-citations will be reduced to the next highest cell in their row and cells will be divided by row total. From the results of this preliminary run decisions can be made as to what further runs are necessary with more of the data.

A secondary category of future work is ancillary to the actual clustering but necessary for full value to be obtained from the results. It consists of two parts, as given below.

(1) Further consideration is necessary of the evaluation, representation and stability of clusters, not necessarily from a statistical viewpoint. This may involve further experimentation with the pilot study data.

(2) As mentioned in section 4.4.1, reduction in the number of columns of the data matrix could be achieved by analysing the data for principal components before

clustering. Principal components of the pilot study data have been obtained but the data has not yet been transformed and clustered. It would be very valuable if a smaller number of principal components than the original number of source journals produces clusters similar to those obtained from the original data.

(3) Consideration of overlapping clusters.

A third category of possible future work consists of subsidiary analyses using other clustering approaches. Such analyses are not essential but might be illuminating while requiring relatively little additional effort. Among these are the following.

(1) Analysis of the pilot study data using the program obtained from Ling (mentioned in section 4.2.3).

(2) Construction by hand of one- and two-step models as proposed by Narin, Carpenter and Berlt (1972) for the ranked (and possibly extra and foreign) source journals. These models can then be compared with the clustering results both of the pilot study and the main data file.

(3) Analysis of the restriction of the data to citations to, as well as from, the source journals by any of the simple clustering methods using a similarity matrix. Different similarity measures could be used. For instance, single- and complete-linkage clusterings can be obtained quite easily by hand from a sorted similarity matrix, or can be very easily programmed if the dendrograms are drawn by hand.

References

AUGUSTSON, J.G. & MINKER, J. (1970). Deriving term relations for a corpus by graph theoretical clusters. Journal of the American Society for Information Science, 1970, 21(2), 101-111.

BORKO, H (1965). A factor analytically derived classification system for psychological reports. Perceptual and Motor Skills, 1965, 29, 393-406.

BORKO, H. & BERNICK, M.D. (1963). Automatic document classification. Journal of the Association for Computing Machinery, 1963, 10, 151-162.

BORKO, H. & BERNICK, M.D. (1964). Automatic document classification part II: additional experiments. Journal of the Association for Computing Machinery. 1964, 11, 138-151.

CARPENTER, M.P. & NARIN F. (1972). Clustering of scientific journals. Chicago, Computer Horizons Inc., 1972.

CUNNINGHAM, K.M. & OGILVIE, J.C. (1972). Evaluation of hierarchical grouping techniques: preliminary study. Computer Journal, 1972, 15(3), 209-213.

DALE, A.G. & DALE, N. (1965). Some clumping experiments for associative document retrieval. American Documentation, 1965, 16(1), 5-9.

DESIGN OF INFORMATION SYSTEMS IN THE SOCIAL SCIENCES (1972). Citation patterns in the social sciences: results of pilot citation study and selection of course journals for main citation study. Bath, Bath University Library, October 1972. (DISISS Working Paper No.5).

DOYLE,L.B. (1962). Indexing and abstracting by association. American Documentation, 1962, 13(4),378-390.

DOYLE,L.B. (1964). Some compromises between word grouping and document grouping. Santa Monica, Calif., System Development Corporation, 1964. AD-440 044

FRIEDMAN,H.P. & RUBIN, J. (1967). On some invariant criteria for grouping data. J. American Statistical Association, 1967, 62(4), 1159-1177.

GITMAN, I (1972).A parameter-free clustering model. Pattern Recognition, 1972,4(4),307-315.

GOTTLIEB, C.C. & KUMAR, S. (1968). Semantic clustering of index terms. Journal of the Association for Computing Machinery, 1968, 15(4),493-513.

HARRISON, P.J. (1968). A method of cluster analysis and some applications. Applied Statistics, 1967, 16(3),226-236

HUBERT, L. (1972). Some extensions of Johnson's hierarchical clustering algorithms. Psychometrika,1972, 37(3),261-274.

JARDINE, N. (1970). Algorithms, methods and models in the simplification of complex data. Computer Journal,1970,13(1), 116-117.

JARDINE, N.& SIBSON, R. (1968). The construction of hierarchic and non-hierarchic classifications. Computer Journal, 1968, 11(2),177-184.

JOHNSON, S.C. (1967). Hierarchical clustering schemes. Psychometrika, 1967, 32(3), 241-254.

KESSLER, M.M. (1963a). An experimental study of bibliographic coupling between technical papers. IEEE Transactions on Information Theory, 1963, 9(1), 49-51.

KESSLER, M.M. (1963b). Bibliographic coupling between scientific papers. American Documentation, 1963, 14(1), 10-25.

KESSLER, M.M. (1965). Comparison of the results of bibliographic coupling with analytic subject indexing. American Documentation, 1965, 16(4), 223-233.

LING, R.F. (1972). On the theory and construction of K-clusters. Computer Journal, 1972, 15(4), 326-332.

NARIN, F. CARPENTER, M. & BERLT, N.C. (1972). Interrelationships of scientific journals. Journal of the American Society for Information Science, 1972, 23(5), 323-331.

PARKER, E.B., PAISLEY, W.J. & GARRETT, R. (1967). Bibliographic citations or unobtrusive measures of scientific communication. Stanford, Stanford University Insitute for Communication Research, 1967.

PREPARATA, F.P. & CHIEN, R.T. (1967). On clustering techniques of citation graphs. Urbana, Ill., University of Illinois, 1967. (Coordinated Science Laboratory Report R-349)

PRICE, N. & SCHIMONOVICH, S. (1968). A clustering experiment: first step towards a computer generated classification scheme. Information Storage and Retrieval, 1968,4(3),271-280

RUBIN, J. (1967). Optimal classification into groups: an approach for solving the taxonomy problem. Theoretical Biology, 1967, 15,103-144.

SALTON, G. (1965) Process in automatic information retrieval. IEEE Spectrum, 1965, 2,90-103.

SALTON, G. (1969) Information science in a Ph.D. computer science program. Communications of the Association for Computing Machinery 1969, 12(2), 111-117.

SCHIMINOVICH, S. (1971). Automatic classification and retrieval of documents by means of a bibliographic pattern discovery algorithm. Information Storage and Retrieval, 1971, 6(6), 417-435.

SCOTT, A.J. & SYMONS, M.J. (1971). Clustering methods based on likelihood ratio criteria. Biometrics, 1971, 27(2),387-397

SHEPARD, R.N., (1962). The analysis of proximities: multidimensional scaling with an unknown distance function, I and II. Psychometrika, 1962, 27(2),125-140 and 219-246.

SIBSON, R. (1973). SLINK: an optimally efficient algorithm for the single-cluster method. Computer Journal, 1973, 16,30-34.

SPARCK JONES, K. (1971). Automatic keyword classification for information retrieval. London, Butterworths, 1971.

STEVENS M.F. (1965). Automatic indexing: a state of the art report. Washington, U.S. Department of Commerce, 1965. (National Bureau of Standards Monograph 91).

STEVENS, M.F., GIULIANO, V.F. & HEILPRIN, L.B. eds. (1965). Statistical association methods for mechnized documentation. Washington, U.S. Department of Commerce, 1965. (National Bureau of Standards, Miscellaneous Publication 269)

STILES, H.F. (1961). The association factor in information retrieval. Journal of the Association for Computing Machinery, 1961, 8, 271-279.

VAN RIJSBERGEN, C.J. (1970). A clustering algorithm. Computer Journal, 1970, 13(1), 113-115.

WILLIAMS, W.T.,LANCE, G.N., DALE, M.B. and CLIFFORD, H.T.,(1971). Controversy concerning the criteria for taxonomic strategies. Computer Journal 1971, 14(2), 162-175.

WOLFF-TERROINE, M. & RIMBERT, D. (1971). Computer-aided automatic generation of a structured documentary language: preliminary study. Journal of Documentation, 1971, 127(2), 111-124.

WORONA,S. (1969). Query clustering in a large document space. In: SALTON, G. (1969). Information storage and retrieval. (Section XV). Ithaca, N.Y.,Cornell University, Department of Computer Science, 1969.

Source journals for the pilot study

Source	Total citations gathered	Average no. of citations /article	Self citations	Self Citations as percentage of total citations	Number of times cited by other source journals
American Anthropologist	504	15	61	12	6
American J. of Sociology	379	9	18	5	23
American Political Science Review	100	14	11	11	6
American Psychologist	128	3	18	14	5
American Sociological Review	37	12	2	5	68
Archives Européenes de Sociologie	176	22	-	-	4
Australian J. of Politics and History	342	20	7	2	-
British J. of Criminology	273	9	6	2	2
British J. of Psychology	490	12	40	8	2
British J. of Sociology	517	21	14	3	5
Economica	291	10	25	8	13
Ecomonic Journal	295	13	29	10	33
Parliamentary Affairs	167	4	3	2	-
Psychologische Forschung	421	52	15	3	3
Revue Economique	401	9	19	4	-
Social Service Quarterly	214	5	4	2	2
Sociological Review	182	16	4	2	5
TOTALS	4918		276		176

APPENDIX B

Pilot Study Citation Data

Cited journals (omitting those cited once only) are listed roughly alphabetically with the number of citations from each of the source journals.

Cited Journal No.	Title	No. of citations from source journals																
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
*001	Acta Physiologica Scand.							2										
002	Acta Psychologica							4	1									
003	Actualité Econ. et Financ.				1							1						
004	Admin. Science Q	1													10			
*005	Adult Education																	2
007	Africa	7		2												3		
008	American Anthropologist	6														2?		
*009	American Antiquity															7		
010	Am.Behavioral Scientist			1	1										1			
011	Am.Economic Review	1								4	10	4						
012	Am.J.Orthopsychiatry				1		4		1								2	
013	Am.J.Psychiatry			1			2										2	
014	Am.J.Psychology							20	4									
015	Am.J.Sociology	7	1	19	3	1		1	1						3	6	1	
016	Am.Mus.Nat.Hist.Anth. papers								1?							2		
017	Am.Pol.Sci. Review												1	11	6			
018	Am. Psychologist				2		19	5										
019	Am.Sociological Review	14	2	19	14	2			13						5	1	2	
*020	Annales(Econ.Soc.Civilis.)											2						
021	Annals	2		2														
022	Annals Math.Statistics							1		1								
023	Annals Am.Acad.Pol. and Soc. Sci.			1						1								
024	Annals N.Y.Acad.Sci.			1				2										
025	Année Sociologique		2			2												

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
*026	Anthropological Papers															2		
*027	Architectural Record																2	
028	Archiv für Gesamte Psych.							1	3									
*029	Archiv für Psych. Nervenkrankheit								3									
*030	Archiv für Psychologie								8									
031	Archiv für Sozialwissenschaft			1						2								
032	Archives de Psychologie							4	1									
033	Archives Eur. de Sociologie		2		2													
034	Aust. and N.Z.J. Sociology	1		2														
*035	Aust.J.Politics & History														7			
036	Australian Outlook														2			
037	Australian Quarterly			1						1								
*038	Behaviour						3											
*039	Betrieb								2									
*040	Biometrics			2														
*041	Birmingham Journal	3																
*042	Brain							3										
043	Br.J.Criminology				2												6	
*044	Br.J.Delinquency																11	
045	Br.J.Educational Psychol.							2				2					1?	
046	Br.J.Medical Psychol.							1									1	
047	Br.J.Psychiatry							1									1	
048	Br.J.Psychology							41							1			
049	Br.J.Psychol.Monograph Supp.							2										
050	Br.J.Sociology	15		3	2										1			
*227	Br.J.Philosophy of Science	9																
051	Bull.Br.Psychological Soc.				1			3										
052	Bull.Oxford Inst.Stat.									1	1							
*053	Bureau of Am.Ethnology Bull.															2		
*054	Cahiers Internationaux											2						
056	Cambridge Journal									1					3			
*057	Canadian J.Corrections								2									
*058	Canadian Hist.Review														2			

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
059	Can.J.Econ. and Pol.Sci.			1							3	1	2				1	
060	Can.J.Psychology							5									1	
*061	Character & Personality								2									
*062	Child Development							8										
*063	China Digest													7				
064	China Weekly Review			1										1				
*065	Chinese Economic J.			2														
066	Colliery Guardian	1			1													
067	Comp.Studies in Soc. & Hist.		1													1		
*068	Current Notes														5			
069	Current Sociology			1	1													
072	Econometrica									7	8							
073	Economic J.	4								28	29	1						
*074	Economic Record										2							
075	Economic Weekly (Bombay)					1				1								
076	Economica			1						26	7	5						
*077	Economie Appliquée											14						
*078	Economist										14							
*079	Educ.Psychol.Measurement							3										
080	Ekonomisk Tidsskrift										2	1						
*081	Electroencephalography, etc.							2										
083	Encounter		1			1												
*084	Erkenntnis								2									
085	Esprit												1		1			
*087	Etudes et Conjoncture											5						
*088	Eugenics Quarterly			2														
*089	Foreign Affairs														3			
*091	Foundation (S.A.)	2																
092	Genetic Psych.Monographs					1		3										
093	Geographical Review			2												1		
*094	Giornale Degli Economisti									3								
*095	Grapholog M.H.								2									
*096	Harper's Magazine			2														
*097	Harvard Educ. Rev.			2														
098	Harvard Law Rev.			1				2										
*099	Historische Zeitschrift					5												
100	Human Organization	4			1										2	1		

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
101	Human Relations	3	1		2												1	
*102	L'Humanité					2												
*103	Industrial Psychotechn.								2									
104	Ind. & Labour Relations Rev.											1			1			
*105	Int.J.Comp.Sociology		2															
106	Int.J. of Social Psychiatry				6												2	
*107	Japanese Psychol.Research							3										
*108	J.Scientific Study of Religion	2																
109	J.Abnormal & Soc.Psychology	3	1	1	1		1	7	7								1	
*110	J.American Folklore															5		
*111	J.Applied Psychology						2											
113	J.Comparative Psychol.						2	2									1	
114	J.Comp. & Physiol.Psychology						2	14										
115	J.Criminal Law & Crim.(SI)	1		1													1	
116	J.Educational Psychology						2	1										
118	J.Experimental Psychology						22		13									
119	J.Genetic Psychology		1					5									1	
*120	J.Marketing									3								
*121	J.Mathematical Psych.							2										
*122	J.Mental Science							5										
*123	J.Negro Education	2																
124	J.Neurology, Neurosurgery, etc							1									1	
125	J.Neurophysiology							6	1									
126	J.Personality							6	7							1		
127	J.Philosophy	2			1													
128	J.Political Economy	1								9	9	3				1		
129	J.Psychology	1						3	2									
130	J.Social Psychology	1					1	3								1		
131	J.Social Issues	2													1		1	
*132	J.Verbal Learning, etc.							11										
*133	J.Acoustical Soc.Am.							3										
*134	J.Am.Medical Ass.			4														
135	J.Am.Statistical Ass.		1				1		1	1								
136	J.Operations Research Soc.Am.	1						1										
137	J.Opt.Society Am.	1								1								
138	J.Royal Statistical Soc.									7	5						1	
*139	J.Siam Society															4		
*140	Kriterion															2		

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
141	Kyklos		1			1												
*142	Lancet																4	
*143	Language							2										
*144	Les Temps Modernes											2						
145	Listener	1				1												1
*146	London & Camb.Econ.Bull.										2							
147	Manchester School	3								1	3							
*148	Medical Economics			8														
*149	Medical J.Australia																2	
151	Monthly Labour Review			1						6								
*152	Nation														2			
153	Nat.Inst.Economic Rev.									1	2	1						
154	Nature	1						15								3	1	
*155	Neue Psychol.Stud.								3									
*156	New Society	3																
157	New Statesman														1			1
159	Occupational Psychology				1			2										
*160	Opinion News			2														
161	Oxford Economic Papers								4?			1						
162	Pacific Sociological Rev.	3		1														
*163	Parliamentary Affairs												3					
*164	Perceptual and Motor Skills							3										
*165	Philosophy of Science	3																
166	Philosophical Studies							2	2									
*168	Population Studies									4								
*169	Praktische Psychologie									3								
170	Proc.Aristotelian Soc.	1			1													
171	Psychiatry				1				1							3	1	
*172	Psychoanal.Stud.Child.						2											
173	Psychol.Arbeiten							1	7									
*174	Psychologia							2										
175	Psychological Bulletin						5	16	2								1	
176	Psychological Monographs	1						6										
*177	Psychological Reports							5										
178	Psychological Review		1					16	13									
*179	Psychologie Rev.								2									
180	Psychologische Forschung							3	15									

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
*181	Psychonomic Science							8										
*182	Psychosomatic Medicine							2										
183	Public Administration												3					
185	Public Opinion Quarterly	2		4											3			
*186	Publications Am.Stat.Ass.			2														
*187	Q.J.Experimental Psychol.							19										
188	Q.J.Economics									1	8	1						
189	Review					1								1				
190	Review of Econ. & Stats.									3	1	3						
191	Rev.Economic Studies										21	4						
*192	Review of Religious Research	4																
193	Revue d'Economie Politique									1		11						
194	Revue Economique											19						
195	Revue Francaise de Sci.Pol.											3	2					
196	Revue Int.de Sociologie			2	1													
*228	Revue Int. du Travail											3						
*197	Revue Pol.et Parliamentaire				3													
*198	Rhodes-Livingstone Papers	3																
*199	Rorschachiana								4									
*200	Scand. J.Psychology							2										
*201	Schrift								2									
202	Science			1	2		1	5								8		
*203	Scottish J.Political Econ.										3							
*204	Sewanee Review		2															
*205	Skand.Arch.Physiol.								2									
207	Social Forces	7		2											11	2		
208	Social Problems	1			3												1	
209	Social Service Quarterly																	4
*210	Social Work	2																
211	Sociological Quarterly			1	2													
212	Sociological Review	4			4											1		
213	Sociometry			2	4													
*214	Sovetskaia Etnografia															6		
215	Studium Gen.							1	2									
216	Survey of Current Business			1						1								
217	Time	1		3														
*218	Trans. Am.Philosophical Soc.															3		
*219	Transports											2						

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
*220	Vierteljahresschrift Wiss.Philos.								3									
*221	Vierteljahresschrift Wiss.Philos. und Soz.								3									
222	World Politics	1	2			2									2			
223	Yale Law Journal			1						1								3
*224	Z.Angewandte Psychologie								23									
*225	Z.Experim.und Ang.Psychol.								6									
*226	Z.Menschenkunde								2									

Notes

1) Titles marked * were omitted from the clustering runs since they were cited by only one source journal. Clearly any reasonable clustering procedure will cluster such journals with their citing source journals and therefore to reduce computer time and storage they can be allocated manually (or by a subsidiary program) after the computer clustering algorithm has been run. A source journal cited by only one journal (usually itself) can be retained for the clustering runs to help this subsidiary allocation.

2) Several transcription and punching errors have been found in this data, but since it was desirable to keep the same data for all runs for comparison purposes these have not yet been corrected. The most important are the omission from the clustering runs as a cited journal of the source journal 14 (cited journal 035) which is cited only by itself, and 4 citations to Oxford Economic Papers by Economica being punched as if from Psychologische Forschung. Cells marked with a ? are known to be in error.

APPENDIX C

Source journals for the main study

JOURNAL TITLE	SOURCE CODE	CLOSS NUMBER	SOURCE LIST CODE	SOURCE	NUMBER OF RECORDS
ABSATZWIRTSCHAFT	ABSATZWIRTS	00042	B	P	*
ACONCAGUA	ACONCAGUA	00083	B	P	21*
ACTA PSYCHIATRICA SCANDINAVICA	ACT PSYC SC	00131	B	I	131*
ACTIVIST	ACTIVIST	00147	B	P	27*
ADMINISTRATIVE SCIENCE QUARTERLY	ADMIN SC Q	00161	F	P	510
ADOLESCENCE	ADOLESCENCE	00164	B	P	291
AFRICA	AFRICA	02634	A	P	186
AFRICAN SCIENTIST	AFRICAN SCI	00217	B	P	*
AGRARWIRTSCHAFT	AGRARWIRTS	00392	B	P	418
ALGEMEN NEDERLANDS TIJDSCHRIFT voor WIJSBEGEERTE en PSYCHOLOGIE	ALG NED PSY	00289	B	P	178*
ALLGEMEINES STATISTISCHES ARCHIV	A STAT ARCH	00296	D	P	122
AMERICAN ANTHROPOLOGIST	AM ANTHROP	00306	A	I	628*
AMERICAN ANTIQUITY	AM ANTIQUIT	04172	A	P	707
AMERICAN ECONOMIC REVIEW	AM ECON R	00324	A	P	1313
AMERICAN FEDERATIONIST	AM FEDER	00326	B	P	66*
AMERICAN JOURNAL OF CORRECTION	AM J CORR	00342	H	P	37*

JOURNAL TITLE	SOURCE CODE	CLOSSS NUMBER	SOURCE LIST CODE	SOURCE	NUMBER OF RECORDS
AMERICAN JOURNAL of ORTHOPSYCHIATRY	AM J ORTHOP	00348	A	I	191*
AMERICAN JOURNAL of PSYCHOLOGY	AM J PSYCHO	00354	A	I	128*
AMERICAN JOURNAL of SOCIOLOGY	AM J SOCIOL	00356	A	I & P	443* + 750
AMERICAN POLITICAL SCIENCE REVIEW	AM POL SC R	00363	A	P	356
AMERICAN PSYCHOLOGIST	AM PSYCHOL	00364	A	I	264*
AMERICAN SOCIOLOGICAL REVIEW	AM SOCIOL R	00373	A	I & P	861* + 537
AMERICAN VOCATIONAL JOURNAL	AM VOC J	00401	B	P	118*
ANALYSIS OF CURRENT DEVELOPMENTS in the SOVIET UNION	ANAL C SOV	00419	B		
ANGLO-NORWEGIAN TRADE JOURNAL	ANGLO-N T J	00431	B	P	22
ANGLO-SOVIET JOURNAL	ANGLO-SOV J	00432	B	P	*
ANNALES:ECONOMIES, SOCIÉTÉS, CIVILISATIONS,	ANN EC S C	00458	D	P	398*
ANNALES DE GEOGRAPHIE	ANN GEOGR	00439	D	P	278*
ARAB VIEWS	ARAB VIEWS	00598	B	P	*
ARCHITECTURE WEST MIDLANDS	A WEST MID	00666	B	P	*
ARCHIV f'r PSYCHOLOGIE	ARCH PSYCHO	04173	A	P	*
ARCHIVES EUROPÉENES DE SOCIOLOGIE	ARCH EU SOC	00697	D	P	268*

JOURNAL TITLE	SOURCE CODE	CLOSSS NO	SOURCE LIST CODE	SOURCE	NUMBER OF RECORDS
ARCHIVES de PSYCHOLOGIE	ARCH DE PSY	00695	A	P	40
ARRIVE	ARRIVE	00723	B	P	19*
BRITISH JOURNAL of CRIMINOLOGY	BR J CRIMIN	00841	?	P	349
BRITISH JOURNAL of EDUCATIONAL PSYCHOLOGY	BR J ED PSY	00842	C	I	*
BRITISH JOURNAL of EDUCATIONAL TECHNOLOGY	BR J ED TEC	00846	C	P	231
BRITISH JOURNAL of PSYCHIATRY	BR J PSYCH1	00858	A	I	700*
BRITISH JOURNAL of SOCIAL WORK	SOC WORK B	02271	C	P	155
BRITISH JOURNAL of SOCIOLOGY	B J SOC	00864	B	P	658
BULLETIN du CENTRE EUROPEEN de la CULTURE	B EU CULTUR	00895	B		
BULLETIN de L'INSTITUT INTERNATIONALE de L'ADMINISTRATION PUBLIQUE	B I ADM PUB	00893	B	P	98
BULLETIN de L'INSTITUT INTERNATIONALE de STATISTIQUE	B I STATIST	03053	D	P	442
CAHIERS FERDINAND de SAUSSURE	CAH FER SAU	00942	B	P	125
CAHIERS INTERNATIONAUX de SOCIOLOGIE	CAH INT SOC	00944	D	P	360
CANADIAN JOURNAL of ECONOMICS	CAN J EC	04876	A	P	444
CANADIAN JOURNAL of PSYCHOLOGY	CAN J PSYCH	00959	A	I	140*

JOURNAL TITLE	SOURCE CODE	CLOSSS NO.	SOURCE LIST ONE	SOURCE	NUMBER OF RECORDS
CHILD DEVELOPMENT	CHILD DEV	00988	A	I	401*
CHINA DIGEST	CHINA DIGES	05408	A		
CRIME and DELIQUENCY	CRIM DELIN	01089	E	P	227*
CRIMINAL LAW REVIEW (LONDON)	CRI LAW R L	04176	E	P	361
DISTRICT COUNCIL REVIEW	DIS COU REV	04174	C	P	*
DIRITTO del LAVORO	DRITTO LAV	02196	B		
EASTERN LIBRARIAN	EAST LIBRAR	02733	B	P	*
ECONOMETRICA	ECONOMETRIC	01142	A	I	303*
ECONOMIA INTERNAZIONALE	ECON INTERN	01143	D	P	261*
ECONOMIC GEOGRAPHY	ECON GEOG	01152	C	P	518
ECONOMIC JOURNAL	ECON J	01155	A	P	505
ECONOMICA	ECONOMICA	01166	A	P	409
ECONOMIE APPLIQUEE	ECON APPLIQ	01171	I	P	423
EDUCATIONAL RESEARCH	EDUC RES	01206	C	P	481
ERGONOMICS	ERGONOMICS	01242	C	I & P	471* + 427

JOURNAL TITLE	SOURCE CODE	CLOSSS NO.	SOURCE LIST CODE	SOURCE	NUMBER OF RECORDS
FILOSOFSKA MISUL	FILOS MISUL	01281	B	P	701*
FINANZ-ARCHIV	FINANZ-ARCH	01283	D	P	246*
FRANÇAIS MODERNE	FRAN MODERN	01296	B	P	247
GEOGRAPHICAL ANALYSIS	GEOG ANALYS	01313	B	P	493
GEOGRAPHICAL JOURNAL	GEOGR J	01316	C	I	77*
GEOGRAPHICAL REVIEW	GEOGR REV	01319	C	P	478*
GEOGRAPHY	GEOGRAPHY	01322	C	P	390
HARVARD BUSINESS REVIEW	HARV BUS RE	01345	C	I & P	26* + 335
HEAD TEACHERS REVIEW	HEAD T REV	01349	B	P	59
HOWARD JOURNAL of PENOLOGY and CRIME PREVENTION	HOWARD J	01378	E	P	95
HUMAN DEVELOPMENT	HUMAN DEV	01380	B	I & P	* + 443
HUMAN ORGANISATION	HUMAN ORGAN	04175	A	P	675
INTERNATIONAL JOURNAL of AMERICAN LINGUISTICS	INT J AM LI	01454	B	P	309

JOURNAL TITLE	SOURCE CODE	CLOSSS NO.	SOURCE LIST CODE	SOURCE	NUMBER OF RECORDS
INTERNATIONAL JOURNAL of PSYCHO-ANALYSIS	INT J PSYCH	01458	C	I & P	315* + 644
INTERNATIONAL JOURNAL of SOCIAL PSYCHIATRY	INT J SOC P	01462	A	P	406
JOURNAL of ABNORMAL PSYCHOLOGY	J ABN PSYCH	01503	A	I	420*
JOURNAL of BUSINESS	J BUSINESS	01513	C	P	437
JOURNAL of CLINICAL PSYCHOLOGY	J CLIN PSYCH	01518	B	I	278*
JOURNAL of COMPARATIVE and PHYSIOLOGICAL PSYCHOLOGY	J COM PHYSL	05059	A	I	1153*
JOURNAL of CONSULTING and CLINICAL PSYCHOLOGY	J CONS CLIN	01528	A	I	419*
JOURNAL of CRIMINAL LAW, CRIMINOLOGY and POLICE SCIENCE	J CRI LA CR	04177	E-	P	252
JOURNAL of CURRICULUM STUDIES	J CURR STUD	01530	C	P	143
JOURNAL of EXPERIMENTAL PSYCHOLOGY	J EXP PSYCH	05079	A	I & P	1232* + 728
JOURNAL of EXPERIMENTAL RESEARCH in PERSONALITY	J EXP RES P	01542	B	P	528
JOURNAL of GENETIC PSYCHOLOGY	J GENET PSY	01545	A	I	*
JOURNAL of the INSTITUTES of EDUCATION at the UNIVERSITIES of NEWCASTLE and DURHAM	J IN EDUC N	01583	B	P	12
JOURNAL of MARKETING RESEARCH	J MARK RES	01553	C	P	257
JOURNAL of PERSONALITY	J PERSONAL	01561	A	I	236*
JOURNAL of POLITICAL ECONOMY	J POL ECON	01563	A	P	1029

JOURNAL TITLE	SOURCE CODE	CLOSSS NUMBER	SOURCE LIST CODE	SOURCE	NUMBER OF RECORDS
JOURNAL of PSYCHOLOGY	J PSYCHOL	01568	A	I	460*
JOURNAL of PSYCHOSOMATIC RESEARCH	J PSYCHOSOM	01569	B	I	758*
JOURNAL of the ROYAL STATISTICAL SOCIETY, SERIES A	J ROY STA A	01591	A	I & P	135* + 313
JOURNAL of SOCIAL PSYCHOLOGY	J SOC PSYCH	01575	A	I	201*
JOURNAL of VERBAL LEARNING and VERBAL BEHAVIOR	J VERB LEAR	01598	A	I & P	560* + 994
KÖLNER ZEITSCHRIFT für SOZIOLOGIE	K Z SOZIOL	04744	D	P	301*
KONTINIST	KONTINIST	04179	D	P	*
KYKLOS	KYKLOS	01609	D	P	235
LIBERAL EDUCATION	LIBERAL EDU	01646	B	P	96
LIBRARY TRENDS	LIB TRENDS	02759	B	P	372
LINGUISTIC INQUIRY	LING ENG	01655	B	P	369
MANAGEMENT TODAY	MANAG T	01701	C	P	178
MANCHESTER SCHOOL of ECONOMIC and SOCIAL STUDIES	MANC SC E S	01707	A	P	244
MONTHLY LABOR REVIEW	MON LAB REV	01794	A	P	291

JOURNAL TITLE	SOURCE CODE	CLOSSS NUMBER	SOURCE LIST CODE	SOURCE	NUMBER OF RECORDS
MUNICIPAL JOURNAL	MUNICIPAL J	04180	C	P	*
MUSIC IN EDUCATION	MUSIC IN ED	01816	B	P	15
NEUEREN SPRACHEN	NEUREN SPR	01887	B	P	718
O and M BULLETIN	OM B	01930	C	P	31
OCCUPATIONAL PSYCHOLOGY	OCCUP PSYCH	01937	B	P	293
OPERATIONAL RESEARCH QUARTERLY	OPERAT R Q	01948	B	I	168*
PARLIAMENTARY AFFAIRS	PARL AFFAIR	01980	C	P	315
PERSONNEL	PERSONNEL	05213	C	I	*
PLANOVOE KHOZIAISTVO	PLANOV KHOZ	04181	D	P	301
PRISON SERVICE JOURNAL	PRIS SERV J	04182	E	P	60*
PROBATION	PROBATION	02045	E	P	27*
PROBLEMI della SICCUREZZA SOCIALE	PROB SI SOC	02047	B	P	168
PSYCHIATRY	PSYCHIATRY	02062	A	I	119*
PSYCHOLOGICAL BULLETIN	PSYCHOL B	02068	A	I & P	1157* + 787

JOURNAL TITLE	SOURCE CODE	CLOSSS NO.	SOURCE LIST CODE	SOURCE	NUMBER OF RECORDS
PSYCHOLOGICAL REVIEW	PSYCHOL R	02967	A	I	287*
PSYCHONOMIC SCIENCE	PSYCHON SCI	05247	A	I	1237*
PUBLIC ADMINISTRATION	PUBL ADMIN	02076	C	P	312
PUBLIC OPINION QUARTERLY	PUBL OPIN Q	02081	A	I	119*
QUARTERLY JOURNAL of ECONOMICS	Q J ECON	02087	A	P	642
QUARTERLY JOURNAL of EXPERIMENTAL PSYCHOLOGY	Q J EXP PSY	04184	A	I	*
RAUMFORSCHUNG and RAUMORDNUNG	RAUMFORSCHU	02112	D	P	263*
REVIEW of ECONOMIC STUDIES	R ECON STUD	02145	F	P	571
REVIEW of ECONOMICS and STATISTICS	R ECON STAT	02144	A	P	699*
REVUE d'ECONOMIE POLITIQUE	R ECON POL	04185	A	P	460
REVUE FRANÇAISE de SCIENCE POLITIQUE	R F SCI POL	04186	D	P	403
REVUE FRANÇAISE de SOCIOLOGIE	R F SOCIOL	02184	D	P	389*
REVUE INTERNATIONALE de CRIMINOLOGIE et de POLICE TECHNIQUE	REV INT CRI	02185	E	P	246
REVUE de PHONETIQUE APPLIQUÉE	REV PHON AP	02174	B	P	79
ROYAL SOCIETY for the PROMOTION of HEALTH - JOURNAL	R SOC HEA J	02213	B	P	363

JOURNAL TITLE	SOURCE CODE	CLOSSS NO.	SOURCE LIST CODE	SOURCE	NUMBER OF RECORDS
SCOTTISH JOURNAL of POLITICAL ECONOMY	SC J POL EC	02242	B	P	282
SMOKELESS AIR	SMOKE AIR	02258	B	P	23*
SOCIAL FORCES	SOCIAL FORC	02262	A	I & P	403* + 669
SOCIAL SERVICE REVIEW	SOC SERV R	02270	C	P	690
SOCIAL WORK (USA)	SOC WORK	04187	C	P	321
SOUTHERN ECONOMIC JOURNAL	SOUT ECON J	02302	B	P	304
SOVIET GEOGRAPHY	SOV GEOGR	02305	B	P	421
SURVEY of CURRENT BUSINESS	S CUR BUSIN	02342	B	P	*
TAXATION	TAXATION	02353	B	P	*
TIERS MONDE	TIERS MONDE	04188	D	P	341*
TIJDSCHRIFT voor ECONOMISCHE en SOCIALE GEOGRAFIE	TIJ EC SO G	02368	J	P	349
TRANSACTIONS of the INSTITUTE of BRITISH GEOGRAPHERS	T IN B GEOG	01423	C	P	584
TRENDS in EDUCATION	TRENDS EDUC	02396	C	P	24
VOPROSY EKONOMIKI	VOPROS EKON	04189	D	P	314

JOURNAL TITLE	SOURCE CODE	CLOSSS NO.	SOURCE LIST CODE	SOURCE	NUMBER OF RECORDS
WEST AFRICAN JOURNAL of EDUCATION	WE AFR J ED	02462	B	P	257
WIRTSCHAFT und STATISTIK	WIRT u STAT	04190	D	P	239
ZEITSCHRIFT für EXPERIMENTELLE und ANGEWANDTE PSYCHOLOGIE	Z E ANG PSY	04191	A	P	405*
ZEITSCHRIFT für PSYCHOLOGIE	Z PSYCHOL	04192	A	P	444
ZEITSCHRIFT für WIRTSCHAFTS und SOZIALWISSENSCHAFTEN	Z WIRTS SOZ	04193	D	P	447*

* Indicates that the data is unknown or that the figure quoted is approximate.

The translation table for the Source List codes is given in table of Appendix

Source: P indicates that the data was collected by the Polytechnic of North London.

I indicates that the data was taken from the I.S.I. tapes.

TC/

APPENDIX D

Criterion used in the SCICON Algorithm for optimizing the division of the points into clusters.

The criterion minimized is the root mean square deviation of points from the centres of the clusters to which they have been allocated.

$$C = \left(\frac{1}{(N-M)} \sum_{i=1}^{N} d_i^2 \right)^{\frac{1}{2}}$$

where N is the total number of points,

M is the number of clusters into which the points are divided,

d_i is the distance from point i to the centre of the cluster to which it is allocated.

Because the average distance of a point from its cluster centre tends to be greater if the number of clusters is reduced, values of the criterion for clusterings into different numbers of clusters are not directly comparable. However consideration of the change in criterion as the number of clusters is reduced will reveal any very distinct optimum number of clusters.

At any stage of the algorithm M is fixed and since C^2 increases monotonically with C, it is sufficient to minimize the sum of squares of deviations of points from their cluster centres

$$C' = \sum_{i=1}^{N} d_i^2$$

The change in this expression when a point is moved from one cluster to another is particularly simple. This fact accounts for the efficiency of the algorithm.

Using Euclidean distance, $d_i^2 = \sum_{j=1}^{J} d_{ij}^2$

where J is the number of dimensions, and d_{ij} is the distance of point i from the cluster centre measured along the jth axis

Hence $$C' = \sum_{i=1}^{N}\left(\sum_{j=1}^{J} d_{ij}^2\right) = \sum_{j=1}^{J}\left(\sum_{i=1}^{N} d_{ij}^2\right)$$

We can therefore conveniently consider the change in a single dimension.

Given a cluster of n elements, the cluster centre is at

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Adding a further element x_s to this cluster moves the centre to

$$\bar{x}' = \frac{1}{n+1}\left\{\sum_{i=1}^{n} x_i + x_s\right\}$$

$$= \frac{n\bar{x} + x_s}{(n+1)}$$

The increase in the contribution of this cluster to C' is

$$\Delta C_+ = \sum_{i=1}^{n}\left\{(x_i - \bar{x}')^2 - (x_i - \bar{x})^2\right\} + (x_s - \bar{x}')^2$$

Now $$\sum_{i=1}^{n}\left\{(x_i - \bar{x}')^2 - (x_i - \bar{x})^2\right\} = n\bar{x}'^2 - 2\bar{x}'\sum_{i=1}^{n} x_i - n\bar{x}^2 + 2\bar{x}\sum_{i=1}^{n} x_i$$

$$= n(\bar{x}'^2 - 2\bar{x}'\bar{x} + \bar{x}^2) = n\,(\bar{x}' - \bar{x})^2$$

$$= \frac{n}{(n+1)^2}\,(n\bar{x} + x_s - (n+1)\bar{x})^2$$

$$= \frac{n(x_s - \bar{x})^2}{(n+1)^2}$$

and $(x_s - \bar{x}')^2 = \frac{1}{(n+1)^2}((n+1)x_s - n\bar{x} - x_s)^2$

$$= \frac{n^2}{(n+1)^2}(x_s - \bar{x})^2$$

$\therefore \Delta C_+ = \frac{n}{(n+1)}(x_s - \bar{x})^2$

Similarly the change in the contribution to C' of a cluster from which an element is removed is

$$\Delta C = -\frac{n}{(n-1)}(x_s - \bar{x})^2$$

... The criterion C' and hence C will be increased by moving x_s from cluster ℓ to cluster m if and only if

$$\frac{n_m}{n_m+1}\sum_{j=1}^{J}(x_{s_j} - \bar{x}_{m_j})^2 < \frac{n_\ell}{n_\ell - 1}\sum_{j=1}^{J}(x_{s_j} - \bar{x}_{\ell_j})^2$$

NOTE that testing for this condition requires only the position of the point in question and those of the two current cluster centres.

APPENDIX E

Clustering results using the SCICON algorithm

The journals in the clusters at the three cluster levels are listed with traces of the clusters to which they were allocated at levels 5, 7, 10, 12, 22. Codes have been allocated to the commonly occurring traces.

APPENDIX B(i) 'Psychology' Cluster

		Level						Trace Code
		3	5	7	10	12	22	
002	Acta Psychologica	2	5	5	5	5	5	A
014	Am.J.Psychology	2	5	5	5	5	5	A
022	Annals Mathematical Statistics	2	5	5	6	6	19	
024	Annals N.Y.Acad of Sciences	2	5	5	5	5	5	A
028	Arch. für Gesamte Psychologie	2	2	7	7	7	7	B
032	Arch. de Psychologie	2	5	5	5	5	5	A
045	Br.J.Educational Psychology	2	5	5	4	4	16	C
046	Br.J.Medical Psychology	2	5	5	4	4	16	C
047	Br.J.Psychiatry	2	5	5	4	4	16	C
048	Br.J.Psychology	2	5	5	5	5	5	A
051	Bull.Br.Psychological Society	2	5	5	5	5	5	A
060	Canadian J.Psychology	2	5	5	5	5	5	A
092	Genetic Psychology Monographs	2	5	5	5	5	5	A
098	Harvard Law Rev	2	5	5	5	5	5	A
109	J.Abnormal & Social Psychiatry	2	2	7	7	7	7	B
113	J.Comparative Psychology	2	5	5	4	12	12	D
114	J.Comparative & Physiological Psych.	2	5	5	5	5	5	A
116	J.Educational Psychology	2	5	5	4	12	12	D
119	J.Genetic Psychology	2	5	5	5	5	5	A
124	J.Neurology, Neurosurgery & Psychiatry	2	5	5	4	4	16	C
125	J.Neurophysiology	2	5	5	5	5	5	A
126	J.Personality	2	2	7	7	7	7	B
129	J.Psychology	2	5	5	7	7	7	
130	J.Social Psychology	2	5	5	5	5	5	A
136	J.Operations Research Society of Am.	2	5	5	5	10	22	
154	Nature	2	5	5	5	5	5	A
159	Occupational Psychology	2	5	5	5	5	5	A

166	Philosophical Studies	2	2	7	7	7	7	B
173	Psychol. Arbeiten	2	2	7	7	7	7	B
175	Psychological Bull.	2	5	5	5	5	5	A
176	Psychological Monographs	2	5	5	5	5	5	A
178	Psychological Rev.	2	2	7	7	7	7	B
180	Psychologishe Forschung	2	2	7	7	7	7	B
215	Studium Gen.	2	2	7	7	7	7	B

APPENDIX E (ii) 'Economics' Cluster

		Level						Trace Code
		3	5	7	10	12	22	
011	Am.Economic Review	3	4	6	8	8	8	A
023	Annals Am.Academy of Political & Social Sci.	3	4	6	6	6	6	B
031	Arch. für Sozial Wissenschaft etc	3	4	6	6	6	6	B
037	Australian Q.	3	4	6	6	6	6	B
052	Bull.Oxford Inst. of Statistics	3	4	6	8	8	20	C
059	Canadian J.	3	1	6	8	8	8	
072	Econometrica	3	4	6	8	8	20	C
073	Economic Journal	3	4	6	8	8	20	C
075	Economic Weekly (Bombay)	3	4	6	6	6	19	
076	Economica	3	4	6	6	6	20	D
080	Ekonomisk Tidsskrift	3	4	6	8	8	8	A
128	J.Political Economy	3	4	6	8	8	20	C
137	J.Opt.Society of Am.	3	3	3	6	10	22	
138	J. Royal Statistical Society	3	4	6	8	8	20	C
147	Manchester School	3	3	3	8	8	22	
151	Monthly Labour Rev.	3	4	6	6	6	6	B
153	National Institute Economic Rev.	3	4	6	8	8	8	
188	Q.J.Economics	3	4	6	8	8	8	A
190	Rev.Economics & Statistics	3	4	6	6	6	20	D
191	Rev.Economic Studies	3	4	6	8	8	8	A
216	Survey of Current Business	3	4	6	6	6	6	B

APPENDIX E(iii) 'Amorphous' Cluster

		Level						Trace code
		3	5	7	10	12	22	
003	Actualite Economique et Financiere	1	1	2	2	9	21	F
004	Administrative Science Q.	1	1	4	3	3	4	D
007	Africa	1	3	3	10	10	10	A
008	Am.Anthropologist	1	3	3	10	10	10	A
010	Am.Behavioral Scientist	1	1	1	9	9	17	B
012	Am.J.Orthopsychiatry	1	1	1	4	12	12	C
013	Am.J.Psychiatry	1	1	1	4	12	14	C
015	Am.J.Sociology	1	1	1	9	11	15	B
016	Am.Museum of Natural History Anthrop.Papers	1	2	7	1	1	13	
017	Am.Political Science Rev.	1	1	4	3	3	11	D
018	Am.Psychologist	1	1	1	4	12	12	C
019	Am.Sociological Rev.	1	1	1	9	11	17	B
021	Annals	1	3	3	9	11	15	
025	Annee Sociologique	1	1	3	1	1	18	G
033	Arch.Europeene de Sociologie	1	1	1	1	9	9	E
034	Australian and New Zealand J.Sociology	1	3	1	9	11	15	
036	Australian Outlook	1	1	4	3	3	4	D
043	Br.J.Criminology	1	1	1	4	4	14	C
050	Br.J.Sociology	1	3	3	10	10	10	A
056	Cambridge Journal	1	1	4	3	3	4	D
064	China Weekly Rev.	1	1	1	9	11	11	B
066	Colliery Guardian	1	3	3	10	10	10	A
067	Comparative Studies in Society & History	1	1	1	1	1	13	E
069	Current Sociology	1	1	1	9	9	9	B
083	Encounter	1	1	3	1	1	18	G
085	Esprit	1	1	4	3	3	3	D
093	Geographical Rev.	1	1	1	9	11	15	B
100	Human Organization	1	3	3	10	10	10	A
101	Human Relations	1	3	3	10	10	10	A
104	Industrial and Labour Relations Rev.	1	1	4	3	3	21	D
106	Int.J.Social Psychiatry	1	1	1	10	9	9	
115	J.Criminal Law and Criminology (SI)	1	3	1	9	11	14	

		Level						Trace code
		3	5	7	10	12	22	
118	J.Experimental Psychology	1	2	7	4	12	12	
127	J.Philosophy	1	3	3	10	10	10	A
131	J.Social Issues	1	3	3	10	10	10	A
135	J.Am.Statistical Assen.	1	1	7	1	12	19	
141	Kyklos	1	1	3	1	1	18	G
145	Listener	1	3	3	1	1	22	
157	New Statesman	1	1	4	3	3	1	D
161	Oxford Economic Papers	1	2	7	7	7	7	
162	Pacific Sociological Rev.	1	3	3	10	10	10	A
163	Parliamentary Affairs	1	1	1	1	1	3	E
170	Proc.Aristotelian Soc.	1	3	3	10	10	10	A
171	Psychiatry	1	1	1	1	1	13	E
185	Public Opinion Q.	1	1	1	9	11	17	B
189	Review	1	1	3	1	1	11	G
193	Rev.d'Economie Politique	1	1	2	2	2	2	F
194	Rev.Economique	1	1	2	2	2	2	F
195	Rev.Francaise de Science Politique	1	1	2	2	2	2	F
196	Rev.Int. de Sociologie	1	1	1	9	11	15	B
202	Science	1	1	1	1	1	13	E
207	Social Forces	1	1	4	3	3	17	D
208	Social Problems	1	1	1	10	9	9	
209	Social Service Q.	1	1	1	1	1	1	E
211	Sociological Q.	1	1	1	9	9	9	B
212	Sociological Rev.	1	3	3	10	10	10	A
213	Sociometry	1	1	1	9	9	9	B
217	Time	1	3	1	9	11	15	
222	World Politics	1	1	3	1	1	18	G
223	Yale Law Journal	1	1	1	4	4	14	C

APPENDIX F

Clusters satisfying Van Rijsbergen's condition

The algorithm suggested by C.J. Van Rijsbergen (1970) was used to identify the tight clusters among the journals in the psychology cluster obtained using the SCICON algorithm.

Ten clusters satisfying the condition that the maximum distance (MAXD) between any pair of points in the cluster should be less than the minimum distance (MIND) of any point in the cluster to any point not in the cluster are as follows :-

			MAXD	MIND
a)	1	ACTA PSYCHOLOGICA (002)	0.0	.47
	7	ARCHIVES DE PSYCHOLOGIE (032)		
b)	5	ANNALS N. Y. ACAD.SCIENCE (024)	0.0	3.93
	15	HARVARD LAW REVIEW (098)		
c)	9	BR.J.MEDICAL PSYCH. (046)	0.0	4.40
	10	BR.J.PSYCHIATRY (047)		
	21	J.NEUROLOGY, NEUROSURGERY AND PSYCH. (124)		
d)	2	AM.J.PSYCHOLOGY (014)	0.34	0.47
	22	J.NEUROPHYSIOLOGY (125)		
e)	1	ACTA PSYCHOLOGICA (002)	.81	1.88
	2	AM.J.PSYCHOLOGY (014)		
	7	ARCHIVES DE PSYCHOLOGIE (032)		
	22	J.NEUROPHYSIOLOGY (125)		
f)	6	ARCHIV FUR GESAMTE PSYCH. (028)	.50	1.88
	34	PSYCHOLOGISCHE FORSCHUNG (180)		
g)	29	PHILOSOPHICAL STUDIES (166)	.82	1.01
	33	PSYCHOLOGICAL REVIEW (178)		

			MAXD	MIND
h)	23	J.PERSONALITY (126)	1.47	2.00
	29	PHILOSOPHICAL STUDIES (166)		
	33	PSYCHOLOGICAL REVIEW (178)		
i)	12	BULL BR. PSYCHOLOGICAL SOC. (051)	1.18	3.00
	28	OCCUPATIONAL PSYCHOLOGY (159)		
j)	3	AMERICAN PSYCHOLOGIST (018)	1.35	3.40
	19	J.EDUCATIONAL PSYCH. (116)		
k)	11	BR.J.PSYCHOLOGY (048)	1.63	1.88
	18	J.COMPARATIVE AND PHYSL.PSYCH (114)		
ℓ)	13	CANADIAN J.PSYCHOLOGY (060)	1.87	2.12
	20	J.GENETIC PSYCHOLOGY (119)		